2003

Block-level discrete cosine transform coefficients for autonomic face recognition

Willie L. Scott, II
Louisiana State University and Agricultural and Mechanical College, wscott@lsu.edu

BLOCK-LEVEL DISCRETE COSINE TRANSFORM COEFFICIENTS FOR AUTONOMIC FACE RECOGNITION

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Interdepartmental Program in
Engineering Science

by
Willie L. Scott, II
B.S., Louisiana State University, 1998
M.S., Louisiana State University, 2001
May 2003

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for blessing me to complete my research; without His divine help, I would have never made it this far. I would like to express my sincere appreciation and thanks to my advisor and major professor, Dr. Subhash Kak, for his constant guidance and valuable comments, without which, this dissertation would not have been successfully completed. My gratitude also goes out to Dr. Jianhua Chen, Dr. Gerald Knapp, and Dr. Suresh Rai, for serving on my defense committee, and special thanks to Dr. Donald Kraft for his willingness to lend a hand. I also wish to thank the members of my family, my caring mom, my encouraging dad, and my cool little brother, for always being there for me and motivating me to stay focused. Last but certainly not least, I want to thank Michele for always being very supportive and reminding me that everything would turn out fine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This dissertation presents a novel method of autonomic face recognition based on the recently proposed biologically plausible network of networks (NoN) model of information processing. The NoN model is based on locally parallel and globally coordinated transformations. In the NoN architecture, the neurons or computational units form distributed networks, which themselves link to form larger networks. In the general case, an *n*-level hierarchy of nested distributed networks is constructed. This models the structures in the cerebral cortex described by Mountcastle and the architecture based on that proposed for information processing by Sutton. In the implementation proposed in the dissertation, the image is processed by a nested family of locally operating networks along with a hierarchically superior network that classifies the information from each of the local networks. The implementation of this approach helps obtain sensitivity to the contrast sensitivity function (CSF) in the middle of the spectrum, as is true for the human vision system. The input images are divided into $N \times N$ blocks to define the local regions of processing. The $N \times N$ two-dimensional Discrete Cosine Transform (DCT), a spatial frequency transform, is used to transform the data into the frequency domain. Thereafter, statistical operators that calculate various functions of spatial frequency in the block are used to produce a block-level DCT coefficient. The image is now transformed into a variable length vector that is trained with respect to the data set. The classification was done by the use of a backpropagation neural network. The proposed method yields excellent results on a benchmark database. The results of the experiments yielded a maximum of 98.5% recognition accuracy and an average of 97.4% recognition accuracy.

An advanced version of the method where the local processing is done on offset blocks has also been developed. This has validated the NoN approach and further research using local processing as well as more advanced global operators is likely to yield even better results.

CHAPTER 1. INTRODUCTION

1.1 Network of Networks Model

The Network of Networks (NoN) model was independently developed by Sutton *et al.* (1988) and Anderson *et al.* (1990). The model is based on two recurrent themes in neurobiology; (1) *groups* of interacting nerve cells, or neurons, encode functional information, and (2) processing occurs simultaneously across different *levels* of neural organization (Guan, *et al*. 1997). As is the case for any model of biological neural networks, the NoN model makes simplifying assumptions about neural activity; however, systems built upon this paradigm have performed very well.

Nested distributed systems denote the core architecture of the model. In the context of describing computational features of the cerebral cortex (Figure l.1a), this organizing principle was first proposed by Mountcastle (1978). Sutton *et al.* (1988) extended this work by providing formal mathematical structure to the proposed theory. Separately, Anderson *et al*. (1990) investigated NoN in the context of signal processing and how the brain sorts out noisy information. In Sutton *et al*., an emphasis on the vertical, or spatial, aspects of the model was explored; in Anderson *et al.*, the horizontal, or temporal, features. Recently, the two approaches have been joined (Sutton and Anderson 1995), Anderson and Sutton 1995). Sutton and Anderson went on to apply their NoN model to the analysis and classification of multiple radar signals that were received simultaneously.

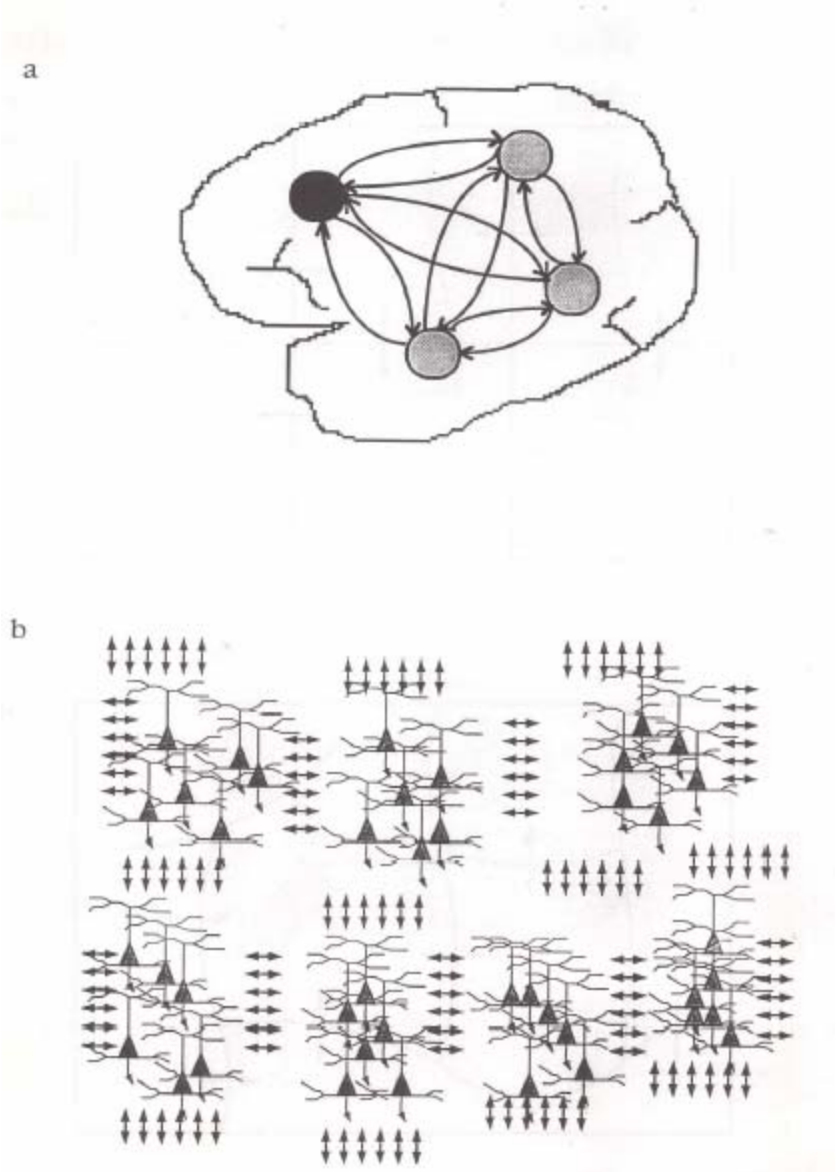


Figure 1.1: Interconnected regions containing networks of networks in the human brain.

Given in Figure 1.1, adapted from Guan *et* al. (1997), is a diagram of how regions of the brain are interconnected, and how the networks inside regions are connected. Figure 1.1a displays a schematic representation of the human brain showing an interconnected network linking four brain regions. In Figure 1.1b, an enlargement of one of the regions in Figure 1.1a (e.g., the black circle). The region contains a network of networks, where seven networks and sample connections are illustrated. In the model,

each network consists of interconnected neurons and is capable of local computations which store and retrieve information in reliable ways. The networks are connected to nearest neighbors to form a larger network. Information encoded within individual local networks is combined and permuted to generate highly complex, time-varying information at a more global level (Guan *et* al. 1997).

Neurons, i.e., computational units, form distributed networks, which themselves link to form larger networks in NoN architectures (Figure 1.1b). Generally speaking, a *n*-level hierarchy of nested distributed networks is constructed (Sutton *et al*. 1988). Neurons maintain their individuality as the networks cluster together to form successively larger networks. In the NoN model, neurons behave as vector entities, which is much different from multi-level networks, which map the activity of a network onto a scalar, and then take the average of scalar networks to map onto the next level of the network.

The formation of clusters and levels among neurons is based on their interconnections (Guan *et al.* 1997). NoN neurons within clusters are more densely connected together than neurons across clusters, and in general, the overall anatomical connectivity is sparse. However, the *functional connectivity* among clusters has rich dynamics, and is due, in part, to temporally correlated activity among neurons (Anderson and Sutton 1995, Vaadia *et al*. 1995).

There are several interesting computational aspects of the NoN model, three of which uniquely characterizing the paradigm (Guan *et al.* 1997):

1. Then NoN model stores and associates information in a hierarchical fashion. High level information consists of nothing more or less than feature combinations among low level information. The storage capacity grows exponentially with the number of levels of clustering, and increased associations within and between levels enhance the performance

of the network. Catastrophic interference, which typically occurs when associations among features or memories becomes too great, is not a significant constraint in the NoN.

2. Individual clusters adapt to learn high-level feature combinations. Network computations can also be directed in specific ways. This is achieved by enhancing or suppressing regions where feature combinations are spatially localized (Anderson *et al*. 1994). Consequently, there is no need for a separate control system to direct the computation.

3. Interaction matrices among individual networks are critical to the computing power of the model (Anderson *et al*. 1994). The interaction terms within a level scale linearly, and not exponentially, with the number of networks. This property is critical for solving real-time problems.

The NoN model suggests that despite enormous diversity in the connection patterns associated with individual neurons, many neural circuits can be subdivided into essentially similar sub-circuits, where each sub circuit contains many types of neurons. This hierarchy is evident in the cerebral cortex, which has long been considered the most complex and elusive of neural circuits (Mountcastle 1978). Cortical sub-circuits are arranged in a nested fashion, with clusters of sub-circuits at the first level coalescing together to form sub circuits at the second level, which cluster to form third-level sub circuits, and so on (See Figure 1.2 adapted from Sutton and Anderson (1995)). This nesting arrangement serves to link different and often widely separated regions of the cortex in a precise but distributed manner. Several physiological responses, such as those occurring in the visual cortex in response to optical stimuli, may be associated with each first-level sub-circuit (Sutton *et al*. 1988).

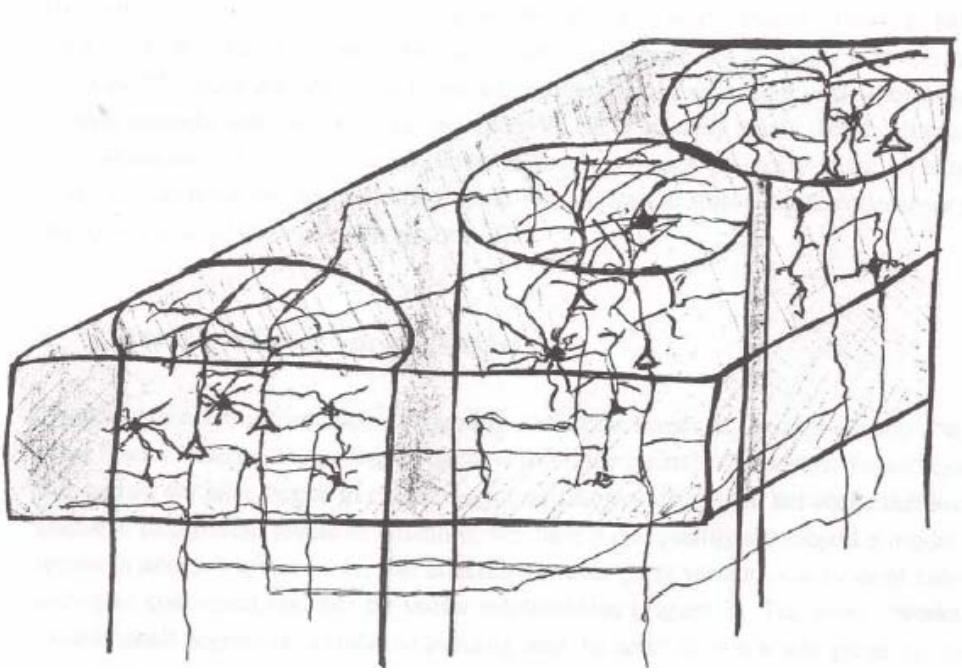


Figure 1.2: Schematic representation of the cerebral cortex. Three networks of intermediate level organization are displayed.

Clustering of first and higher-level sub-circuits form the boundary relations, and are associated with successive levels of complex cortical behavior that integrate and modify responses at lower levels. This simplifying viewpoint of neurobiology of the cerebral cortex captures an important design strategy that has been largely overlooked in mathematical studies of neural circuits.

In the model, first-level clusters are linked together by subsets of elements, referred to as projection elements, to form second-level clusters. Other subsets of projection elements join second-level clusters to produce third-level clusters. Continuing in this manner, an *r*-level hierarchy of nested clusters of neural elements is constructed (Figure 1.3). In the NoN model, subsets of projection elements are assumed to be mutually exclusive for modeling purposes. In this way, $r - 1$ populations of projection elements are identified within each first-level cluster, such that each population establishes connections with elements in its own cluster and with elements in other

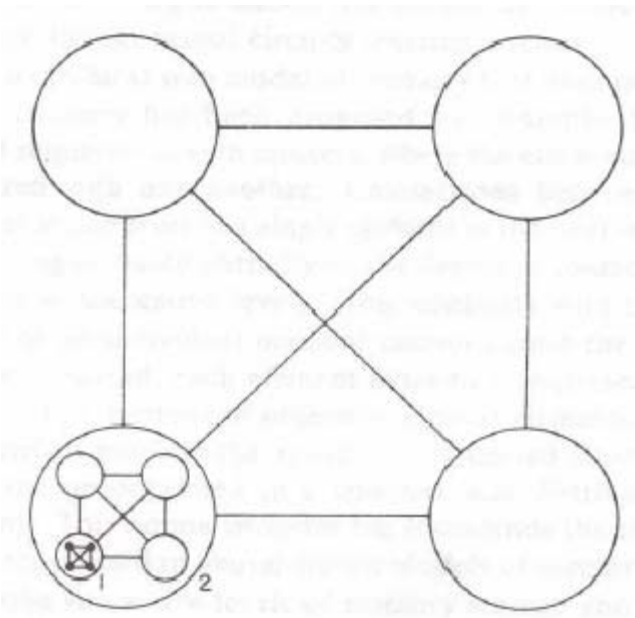


Figure 1.3: Hierarchy of neural clusters demonstrating levels of organization. Clusters at the first level contain neural elements (small dots) which are fully interconnected with each other, as shown in cluster 1. Certain linkages between first-level clusters form second-level clusters, such as cluster 2.

clusters nested together at a unique level (Sutton *et al*. 1988).

In the model, each first-level cluster has several first-level memory states and the second and higher-level memory states consist of sets of correlated first-level memory states. In this context, the correlations are determined by the details of the intercluster connections. Memory is hierarchical in the sense that lower-level memory states combine to create emergent and possibly degenerate higher-level memory states; higher-level states, in turn, categorize lower-level states (Sutton *et al*. 1988).

In the NoN model, hierarchical memory due to nesting of neural clusters arises from design principles found in complex neural circuits. Elements in each cluster are connected with one another, and the activity of an individual element does not depict the collective activity of many other elements (Figure 1.4). Here, each element is taken to represent a single neuron-like entity and the graded projections of otherwise similar elements are utilized to construct a hierarchical circuit model. The result is that nested

clusters and their associated memory properties are organized in a complex and distributed manner at each level of the hierarchy. This notion includes but transcends the

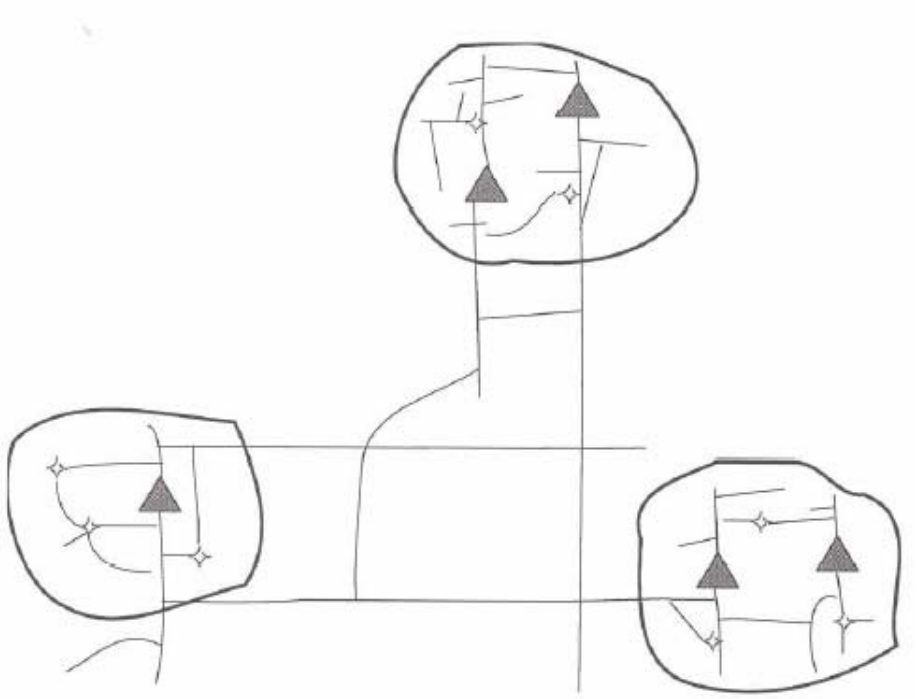


Figure 1.4: A network of three sub-networks. Vector connectivities among the model neurons are maintained within and between the sub-networks. Adapted from Sutton and Anderson (1995).

traditional sequential arrangement of clusters utilized in neural circuit models of memory, describing successive levels of memory storage and recall (Sutton *et al*. 1988).

The NoN method has yet been used, with excellent results, for image regularization (Guan *et al.* 1997), a processing technique which attempts to alleviate the degradations and provide clear and noise-free images, and robotics by Sutton and Anderson, and their associates. This dissertation extends the application of this biologically plausible approach to the problem of face recognition.

1.2 Hierarchical Processing of Face Images in the Human Vision System

The application of the NoN approach is well suited for the problem of face recognition, due to analogous manner in which faces are processed in the human vision

system (HVS). In the HVS, both global and local features are used for face recognition in a hierarchical fashion, the local features providing a finer classification system for facial recognition (Hay and Young 1982). In this section we give more information concerning the HVS and related biological processes.

1.2.1 Human Capacity for Face Recognition

The study of the relationship between the human capacity for face recognition and machine face recognition is not a novel concept. Much work has been done in this area, with the goal of better modeling the human ability for face recognition. There is evidence to suggest that the human capacity for face recognition is a dedicated process, not merely an application of the general object recognition process (Ellis 1986). This may have encouraged the views that artificial face recognition systems should also be face-specific. Determining which features humans use for face recognition has been subject to much debate, and the result of the related studies has been used in the algorithm design of some face recognition systems.

When considering human recognition of faces, the most difficult faces to recognize are those faces which are considered neither attractive nor unattractive by the observer. This gives support to the theories suggesting that distinctive faces are more easily recognized than typical ones (Baron 1981). Young children typically recognize unfamiliar faces using unrelated cues, such as glasses, clothes, hats, and hair style. By age twelve, these paraphernalia are usually reliably ignored (Carey *et al*. 1980). Psychosocial conditions also affect the ability to recognize faces. Humans may encode an "average" face; these averages may be different for different races, and recognition may suffer from prejudice or unfamiliarity with the class of faces from another race (Goldstein 1979b) or gender (e.g., Goldstein 1979a, Weng and Swets 1999).

1.2.2 Spatial Vision

The human ability to recognize faces efficiently and accurately, as well as other forms of visual stimuli, is facilitated by *spatial vision*. Spatial vision is considered perhaps the most important aspect of human sight, which is defined as the ability to see external objects and discriminate their different shapes. In everyday life, spatial vision is necessary in many instances, while many individuals take for granted this capability. The acts of reading road signs or instructions on a printed document, urban driving in rain or fog, crossing a busy street are all facilitated by the HVS's capacity for spatial vision. The familiarity of these demands blinds us to the remarkable discrimination and reliability of everyday spatial vision, such as our ability to recognize one out of possibly thousands of faces stored in memory (Regan 1991). Vision research has shown that there is a substantial amount of neural processing that underlies spatial vision. Research in the area of spatial vision endeavors to understand functional principles and physiological operations of a rather substantial part of the human brain, a structure containing some 10^{10} neurons, each of which receives up to thousands of connections from other neurons, and whose state continually changes (Regan 1991).

1.2.3 The Contrast Sensitivity Function

In early spatial vision research, the concept of spatial frequency orientation selectivity was shown to be correct in the HVS. This important concept is captured by the contrast sensitivity function (CSF), a transfer characteristic used to model the HVS as a measure of the organism's response to various spatial frequencies. Cannon (1979) conducted psychophysical experiments that presented different spatial frequencies at a variety of fixed contrasts and required the subjects to give a contrast-dependent response. Cannon typically chose the response to be the absolute contrast threshold for the

detection of a pattern. The resulting function was termed the CSF, contrast sensitivity being expressed as the reciprocal of the contrast required for detection.

Given in Figure 1.5 is a typical CSF for a normal human observer, obtained by utilizing images with luminance-varying gratings with central fixation at a stimulus light level. The definition of gratings will be given in the following passages. The most important notion here is that the highest sensitivity is in the mid spatial frequency range, with a drop in sensitivity to high spatial frequencies, and a gentler but still pronounced loss in sensitivity at low spatial frequencies as well, when frequency is plotted on a logarithmic scale. The CSF displays that the HVS utilizes information from both ends of the spatial frequency spectrum, but focuses on the information in the middle of the spectrum.

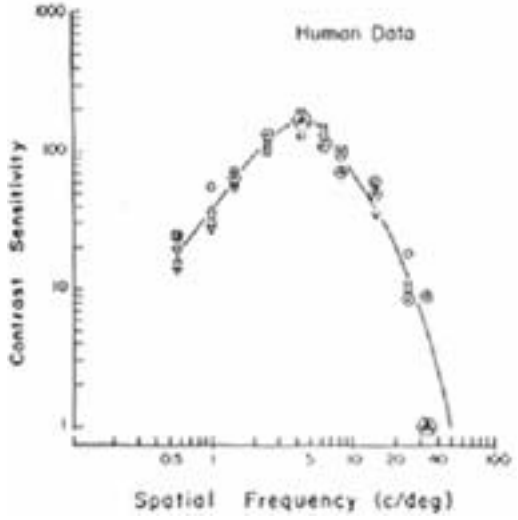


Figure 1.5: A typical CSF for a normal human observer. The units of spatial frequency are cycles per degree.

In recent times, the CSF has become a standard index of the visual system's sensitivity to pattern information of various sorts. The CSF is determined by measuring an observer's sensitivity to sine wave gratings of different spatial frequencies. A sine

wave grating is a pattern composed of regularly spaced light and dark stripes (see Figure 1.6).

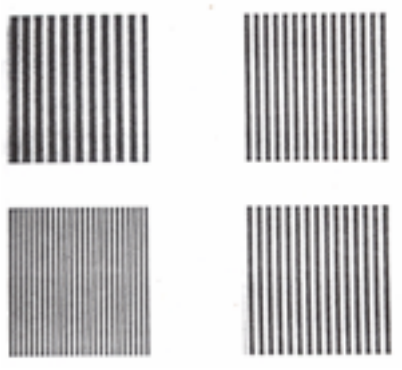


Figure 1.6: Example sine wave gratings.

Sine wave gratings are specified by three parameters: (1) spatial frequency, the number of light stripes per degree of visual angle; (2) orientation; and (3) contrast. The reason sine wave gratings are used derives from Fourier's theorem and linear systems analysis. Fourier's theorem implies that any two-dimensional pattern can be exactly described by combining a set of sine wave gratings of various spatial frequencies, orientations, and contrasts. Cornsweet (1970) and Banks and Salapatek (1981) utilized this technique to describe spatial frequency information in terms of the coarseness of pattern information. Their research was able to successfully display that low spatial frequencies corresponded to coarse pattern information in images, such as the outline shape. High spatial frequencies, on the other hand, corresponded to fine pattern information such as the texture of a surface.

1.2.4 Channels

Cornsweet (1970) and Banks and Salapatek (1981) successfully displayed that spatial vision in humans employed spatial frequency orientation selectivity, that different and important forms of information were present in high and low spatial frequency sub-

bands, and that information corresponding to these sub-bands were used in a hierarchical fashion. In this section, we will discuss in detail how spatial frequency orientation selectivity is facilitated in the HVS; namely, via *channels,* which are to be defined.

The HVS is modeled as a system characterized by a response relating the output to input stimuli. Here, models are validated by psychophysical experiments in which human subjects are asked to assess visibility of stimuli. Electro-physiological experiments performed on cells of the primary visual cortex (area V1 of Figure 1.7) have shown that the response of such neurons is tuned to a band limited portion of the frequency domain (De Valois and De Valois 1988). These results have been confirmed by additional psychophysical experiments in Daugman, (1984), giving evidence that the brain decomposes visual stimuli into so-called *perceptual channels* (Daugman 1984) that are bands in spatial frequency orientation and temporal frequency. Each channel can thus be seen as the output of a filter which is characterized by a response tuned to a specific spatial frequency orientation and temporal frequency.

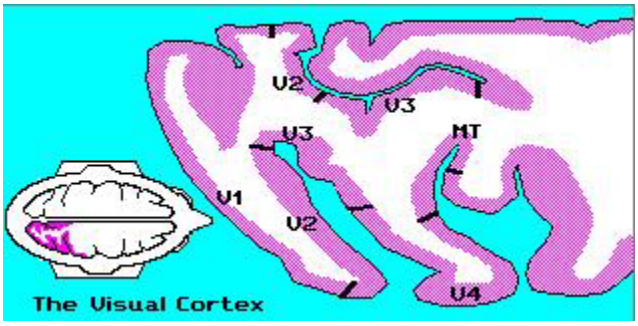


Figure 1.7: The visual cortex. The visual cortex 'computes' sight using the data collected by the eyes. Adapted from Heuther (2002).

Recall the concept of the CSF, a transfer characteristic representing spatial vision in the HVS, which demonstrates that the vision system utilizes information from both ends of the spatial frequency spectrum. There is now a considerable body of psychophysical, physiological, and anatomical evidence supporting that the CSF

represents the envelope of narrowly tuned channels. Research has shown that a mature visual system possesses mechanisms that analyze visual inputs into bands of spatial frequency. It is important to note that this analysis appears to be very important to several visual capabilities. The notion of a channel refers to a filtering mechanism, something which passes some, but not all, of the information that may impinge upon it. A recognized biological definition of a channel is that it is composed of all those cells with *receptive fields*[1] (RFs) that are identical in every respect save retinal location, each with a certain RF periodicity which gives it a particular spatial frequency bandwidth (Graham 1990). Generally speaking, spatial frequency filters may transmit whatever input is presented above a particular frequency (high-pass), below a particular frequency (low-pass), or within a restricted frequency region, with rejection at both end (band-pass). The concept of spatial frequency orientation selectivity in the HVS has two important facets. Firstly, that spatial frequency filters are responsive to only some fraction of the total range encompassed by the organism's vision system. The second important concept is parallel processing of information. A pattern within a given limited spatial region is not analyzed on the basis of the point-by-point luminance level, but rather is simultaneously analyzed by multiple spatial frequency filters. Here, spatial filters are only responsive if the spatial pattern contains energy within its particular spatial frequency band (De Valois and De Valois 1988).

The HVS has clearly been shown to possess parallel pathways (i.e. channels) each specialized to carry information about particular types of stimuli. Moreover, electrophysiological investigations have shown that different sorts of information from the same location in the visual field are signaled by different neurons. More recent

[1] The term "receptive field" refers to the set of one or more retinal receptors which transmit impulses to a given cell in the nervous system.

evidence suggests that the stimulus preferences of cortical neurons can be well described in spatial frequency terms; that is, the preference of each neuron is limited to a range of orientations and a band of spatial frequencies (De Valois and De Valois 1988). So one cell might respond maximally to contours that are vertical and low in spatial frequency and another to contours that are horizontal and high in frequency (e.g., Uhr *et al*. 1962, Campbell *et al*. 1969, Movshon *et al*. 1978, and Albrecht *et al*. 1980). This evidence implies that the responses of single cortical neurons convey orientation and spatial frequency information from a particular region in the visual field (De Valois and De Valois 1988). Different channels appear to be tuned to different orientations and spatial frequency bands (e.g., Campbell and Robson 1968, Blakemore and Campbell 1969, Graham and Nachmias 1971, Braddick *et al*. 1978).

Neurons in the primary visual cortex (V1) have been shown to respond preferentially to stimuli with distinct orientations and spatial frequencies. The method by which the HVS processes information in the spatial frequency spectrum has been researched by Marr (1982) and his colleagues. Marr successfully displayed that channels are used to segregate pattern information early in visual processing in a way that will be useful upstream. Specifically, this segregated pattern information is combined in a hierarchical manner. The research ascribes a relatively peripheral role to spatial-frequency-selective channels. For illustration, Marr examined the problem of visual scene analysis, specifically, distinguishing intensity differences in a scene that are caused by shadows or highlights from intensity differences caused by true object boundaries. Marr displayed that comparison of low spatial frequency (low-pass) and high spatial frequency (high-pass) representations of a scene may allow the visual system to distinguish contours caused by shadows and highlights from those created by object boundaries or

discontinuities. Figure 1.8 (Banks *et al*. 1985) illustrates this phenomenon. Three versions of a photograph are shown. The first (Fig. 2.5A) is the original, unfiltered picture of the face. The second (Fig. 2.5B) is a low-pass version: medium and high spatial frequencies have been filtered out so only very low frequencies remain. The third (Fig. 2.5C) is a high-pass version: only fairly high frequencies are represented. Note that distinct object boundaries (like the contour formed by the cheek and the background) appear in all three images whereas intensity gradients due to shadows (e.g., the region between the woman's right eye and her nose) and highlights (e.g., the bright spot on the chin) appear in only the original and low-pass images. Marr and colleagues hypothesized that the HVS may correlate such filtered representations to distinguish intensity gradients due to object boundaries or discontinuities from those due to lighting conditions.

Figure 1.8: Three photographs of a face. Top-left (1.8A) is the original, unfiltered photograph of the face. Top-right (1.8B) is a filtered photograph of the same face. Medium and high spatial frequencies have been filtered out so only low spatial frequencies remain. Bottom (1.8C) is another filtered photograph of the face. Here, low frequencies have been filtered out so only higher spatial frequencies remain.

Marr's research had a large impact on vision because it provided a model of how the visual system might be analyzing spatial patterns which was quite different from mechanisms previous works had considered. In this new model, the visual system was considered to be operating not purely in the space domain, analyzing the amounts of light at different points in space or edges where the amount of light abruptly changes, but rather operating partially in the spatial frequency domain, responding to patterns on the basis of underlying frequency content.

1.2.5 Interaction of Channels in Human Vision

Although the definition of channels involves their mutual independence in some respect, the channels show a level of interaction in the HVS. When tested in vision research (Henning *et al*. 1975), channels were shown to not only respond to a particular frequency band, but also to the corresponding spatial periodicity in components well outside of that band.

More specifically, Blakemore and Sutton (1969), and Blakemore and Carpenter (1970) have demonstrated the existence of spatial frequency channels, and that they are involved in the detection of contrast. The authors proposed that if there are distinct channels responding to different spatial frequencies, it becomes very plausible to reason that the hierarchical distribution of activity among these channels is the means by which the spatial frequency content of an image is coded in the visual system. In addition to distribution of work among the channels, there is significant interaction between the channels, which serves as the major vehicle for vision.

Hierarchical interactivity between the channels is necessary for human vision. Much research in the area of human vision has explored the consequences of filtering out higher spatial frequency components of the visual input (e.g., Kabrisky *et al*. 1970,

Ginsburg 1971, Carl and Hall 1972). The resulting low-pass filtered image typically loses much of the detail found in the input stimuli. Ginsburg (e.g., Ginsburg 1971, Ginsburg *et al*. 1976) emphasizes many of the classic "Gestalt" organizational properties

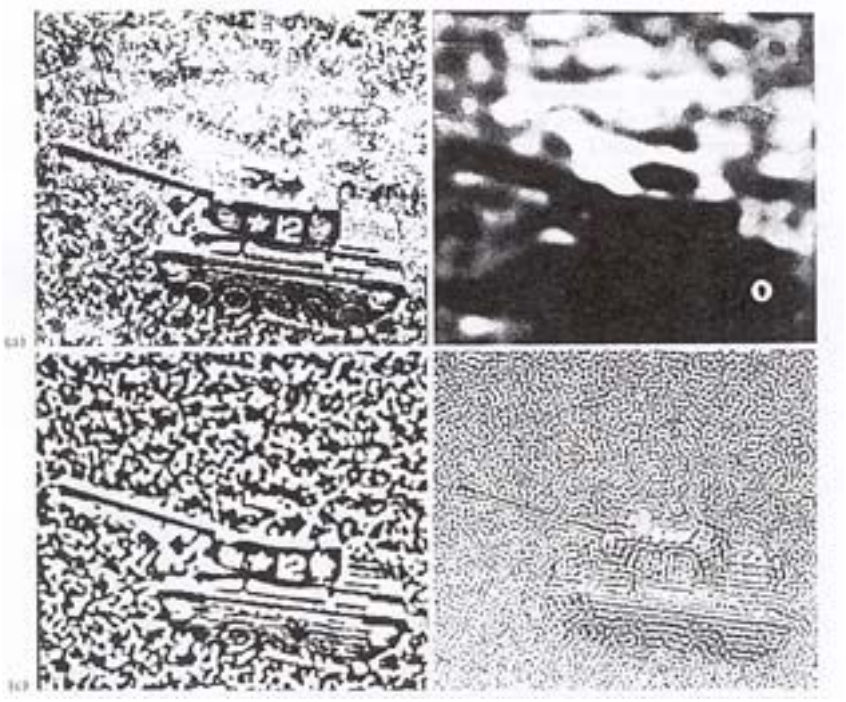


Figure 1.9: A simulation of different types of information present in spatial frequency bands. Top-left is the original photographic image. Top-right is the same image filtered to pass low spatial frequencies only. Bottom-left is filtered to pass middle spatial frequencies only. Bottom-right is filtered to pass high spatial frequencies.

of patterns. It is suggested that this phenomena, when observed in image recognition, may depend on the processing of a low-pass filtered section of the frequency spectrum. However, Ginsburg successfully displayed that attention must be distributed in a variable way among the channels dealing with different parts of the frequency spectrum. For certain perceptual judgments, information from the low frequency channels is used; for others, information from the high frequency channels (Ginsburg 1971). Given in Figure 1.9 (Braddick *et al*. 1978) is a simulation of the different types of information which could be extracted, as attention is shifted between different bands of frequency. Note that

the global form present in the image is transmitted most prominently in the low frequency band, but for information about details such as the number, high spatial frequencies must be examined.

1.3 Image Transforms

The face recognition problem also involves the transformation of faces into the frequency domain for processing. This transformation needs to be coupled together with the NoN approach. In the rest of the introduction, we summarize important image transforms and relevant related issues.

We begin with a discussion of linear transforms. A complete linear transform represents a signal as a weighted sum of *basis* functions. That is, a signal, *f(x)*, is represented as a sum over an indexed collection of functions, $g_i(x)$:

$$f(x) = \sum_i y_i g_i(x)$$

where the y_i are the transform coefficients. These coefficients are computed from the signal by projecting onto a set of functions that we call the *projection functions*[2], $h_i(x)$:

$$y_i = \int_0^{2\pi} dx \; h_i(x) f(x)$$

The equation above is used to compute the transform of the signal. The original signal may be recovered from the coefficients, y_i, and the basis functions $g_i(x)$, by using the equation expression for *f(x)*.

Many image transforms are based on convolution operations. In transforms based on convolution operations, the projection functions are shifted copies of the (reversed)

[2] In general, the projection functions P_i^n, which take n arguments and return their i^{th} argument, are given as: $P_i^n\left(\vec{x}^n\right) = x_i$ for any $1 \le i \le n$ and any $\vec{x}^n \in \mathbb{R}^n$.

convolution kernels. From the definition of convolution, the projection functions corresponding to a single convolution and sampling stage are:

$$h_i(x) = h(i\Delta_x - x).$$

Here, the subscript *i* indexes the spatial position of its associated basis function, and the sample spacing is denoted by Δ_x. The basis functions, given by $h_i(x)$, are said to be *linearly independent* if there is no linear combination of them that is zero for all *x*. If the set of basis functions is not linearly independent, then the transform is said to be *overcomplete*. In this case, the corresponding projection functions are not unique. We will refer to a transform for which $h_i(x) = g_i(x)$ as *self-inverting*. If the transform is self-inverting and the basis functions are linearly independent, then the transform is described as *orthonormal*[3].

In selecting image transforms for use in autonomic recognition systems, often, one is concerned with the properties of the individual basis and/or projection functions. In particular, recognition schemes based on image signal analysis often require the use of functions that are spatially localized (sometimes referred to as "tuned"). The property of localization implies a localization measure: for example, a function that is localized in space might have a restricted region of support, or a finite variance (second moment). Similarly, one can consider functions that are localized in the Fourier domain. A function that has a localized Fourier spectrum is tuned for a particular range of scales. Image frequency transforms like the Fourier transform, in a rudimentary sense, may be

[3] A subset $\{v_1, ..., v_k\}$ of a vector space V, with inner product $\langle , \rangle$, is called orthonormal if $\langle v_i, v_j \rangle = 0$ when $i \neq j$. That is, the vectors are mutually perpendicular. Moreover, they are all required to have length one: $\langle v_i, v_i \rangle = 1$.

considered as a sum of spatially sinusoidal components of different frequencies. The frequency spectrum indicates the magnitudes of the spatial frequencies contained in an image. Figure 1.10 displays a general two-dimensional image frequency transform.

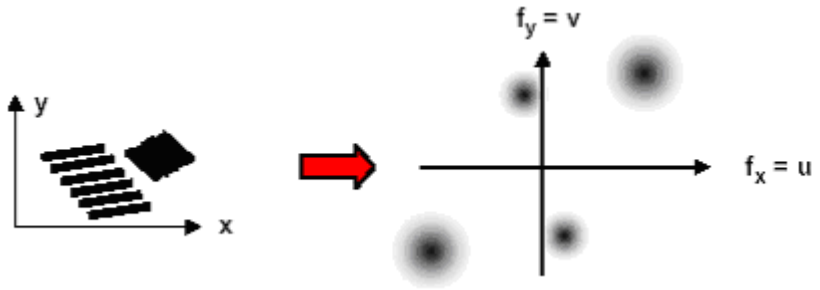


Figure 1.10: A general frequency transform.

A function that is simultaneously tuned for several parameters is said to be jointly localized in those parameters. The concept of joint localization in space and spatial frequency was introduced by Gabor (1946). He derived a lower bound for a particular measure of joint localization, and determined that unique set of function (products of sinusoids and Gaussians) that achieves this limit.

Another set of image transformations, while not as widely used in recognition systems, are known as wavelet transformations (e.g., Grossman and Morlet 1984, Daubechies 1988). In an independent context, mathematicians developed a form of continuous function representation called wavelets. The defining characteristic of a wavelet transform is that the basis functions are translations and dilations of a common kernel (Strang 1989). Therefore, the wavelet transform is denoted as a scale decomposition, and in most literature, the term wavelet is usually assumed to refer to an orthonormal basis set (Chen and Pratt 1984).

1.3.1 Discrete Fourier Transform

In most instances, input images to an autonomic image recognition system are compressed to reduce computational and memory costs. The Discrete Fourier Transform (DFT) performs compression by computing the image representation as a sum of sinusoidals. The DFT is given by $G_{uv} = \mathcal{F}\{g_{mn}\}$ where:

$$G_{uv} = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} g_{mn} e^{-2\pi\left(\frac{mu}{M}+\frac{nv}{N}\right)}$$

and the inverse DFT is given as $g_{mn} = \mathcal{F}^{-1}\{G_{uv}\}$ as

$$g_{mn} = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} G_{uv} e^{2\pi\left(\frac{mu}{M}+\frac{nv}{N}\right)}.$$

In general, the Fourier transform is a complex function with a real and an imaginary part:

$$G_{uv} = R_{uv} + \mathrm{i}\, I_{uv}$$

having frequency spectrum or amplitude spectrum given as

$$\left|G_{uv}\right| = \sqrt{R_{uv}^2 + I_{uv}^2}.$$

The power spectrum or spectral density is defined as

$$P_{uv} = \left|G_{uv}\right|^2 = R_{uv}^2 + I_{uv}^2$$

and phase spectrum given as

$$\Phi_{uv} = \tan^{-1}\left(\frac{I_{uv}}{R_{uv}}\right).$$

In Section 1.3 we discuss the image transform of interest for this dissertation, namely, the Discrete Cosine Transform (DCT). The DCT is one of the most commonly used transforms in practice for image compression, even more so than the Discrete

Fourier transform. This is because the DFT assumes periodicity which is not necessarily true in images. In particular, to represent a linear function over a region requires many large amplitude, high frequency components in a DFT. This is because the periodicity assumption will view the function as a sawtooth which is highly discontinuous at the teeth requiring the high frequency components. The DCT does not assume periodicity and will require much lower amplitude high frequency components. The DCT also does not require a phase component, which is typically represented using complex numbers in the DFT.

1.3.2 Karhunen-Loeve Transform

Much research in the area of image recognition focuses on the Karhunen-Loeve Transform (Karhunen 1947). The Karhunen-Loeve Transform (KLT) is a rotation transformation that aligns the data with the eigenvectors, and decorrelates the input image data. Here, the transformed image may make evident features not discernable in the original data, or alternatively, possibly preserve the essential information content of the image, for a given application with a reduced number of the transformed dimensions. The KLT develops a new coordinate system in the multispectral vector space, in which the data can be represented without correlation as defined by:

$$\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{x}$$

where $\boldsymbol{Y}$ is a new coordinate system, $\boldsymbol{G}$ is a linear transformation of the original coordinates that is the transposed matrix of eigenvector of the pixel data's covariance in $\boldsymbol{x}$ space, and $\boldsymbol{x}$ is an original coordinate system. By the equation for $\boldsymbol{Y}$, we can get the principal components and choose the first principal component from this transformation that seems to be the best representation of the input image.

The KLT, a linear transform, takes the basis functions from the statistics of the signal. The KLT transform has been researched extensively for use in recognition systems because it is an optimal transform in the sense of *energy compaction*, i.e., it places as much energy as possible in as few coefficients as possible. The KLT is also called *Principal Component Analysis* (PCA), and is sometimes referred to as the *Singular Value Decomposition* (SVD) in literature. The transform is generally not separable, and thus the full matrix multiplication must be performed:

$$\boldsymbol{X} = U^T\boldsymbol{x}, \quad \boldsymbol{x} = U\boldsymbol{X},$$

Where U is the basis for the transform, estimated from a number of x_i, $i \in [0,1,\ldots,k]$,

Where

$$U\sum V^T = [x1, x2, ..., xk] = A \Rightarrow U = eigvec(AA^T).$$

The KLT has been used in vision research for several reasons. As stated previously, the KLT transform is optimal in terms of energy compaction; however, the Discrete Cosine Transform closely approximates the KLT with the benefit of being much less computationally expensive. The KLT identifies the dependence structure behind a

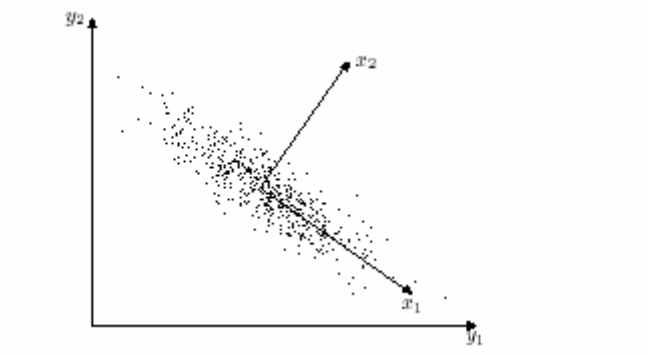


Figure 1.11: Example of a KLT transform.

multivariate stochastic observation (in our context, an input pixel image) in order to obtain a compact description of it (Diamantaras *et al*. 1996). When the observed variables

have a non-zero correlation, the dimensionality of the data space does not represent the number of independent variables, or hidden factors, that are really needed to describe the data. The more correlated the observed variables, the smaller the number of independent variables that can adequately describe them. Given in Figure 1.11 (De Baker 2002) is an example of a transformation from a two dimensional dataset in axes, (y_1, y_2), to its decorrelated KLT representation (x_1, x_2).

1.3.3 The Discrete Cosine Transform

A discrete cosine transform (DCT) is a sinusoidal unitary transform. The DCT has been used in digital signal and image processing and particularly in transform coding systems for data compression/decompression. This type of frequency transform is real, orthogonal, and separable, and algorithms for its computation have proved to be computationally efficient. In fact, the DCT has been employed as the main processing tool for data compression/decompression in international image and video coding standards (Rao and Hwang 1996).

The family of DCTs has been described by Wang and Hunt (1985). The set consists of 8 versions of DCT, where each transform is identified as even or odd and of type I, II, III, and IV. All present digital signal and image processing applications (mainly transform coding and digital filtering of signals) involve only even types of the DCT. As this is the case, the discussion in this dissertation is restricted to four even types of DCT.

In subsequent sections, N is assumed to be an integer power of 2, i.e., $N = 2^m$. A subscript of a matrix denotes its order, while a superscript denotes the version number. Four even types of DCT in the matrix form are defined as (Wang and Hunt 1985):

$$\text{DCT} - I: \qquad \left[C_{N+1}^{I}\right]_{nk} = \sqrt{\frac{2}{N}}\left[\varepsilon_n \varepsilon_k \cos\frac{\pi nk}{N}\right]$$
$$n, k = 0, 1, \ldots, N - 2$$

$$\text{DCT} - II: \qquad \left[C_{N}^{II}\right]_{nk} = \sqrt{\frac{2}{N}}\left[\varepsilon_k \cos\frac{\pi(2n+1)k}{2N}\right]$$
$$n, k = 0, 1, \ldots, N - 1$$

$$\text{DCT} - III: \qquad \left[C_{N}^{III}\right]_{nk} = \sqrt{\frac{2}{N}}\left[\varepsilon_n \cos\frac{\pi(2n+1)n}{2N}\right]$$
$$n, k = 0, 1, \ldots, N - 1$$

$$\text{DCT} - IV: \qquad \left[C_{N}^{IV}\right]_{nk} = \sqrt{\frac{2}{N}}\left[\cos\frac{\pi(2n+1)(2k+1)}{4N}\right]$$
$$n, k = 0, 1, \ldots, N - 1,$$

where

$$\varepsilon_p = \begin{cases} \dfrac{1}{\sqrt{2}} & p = 0 \text{ or } p = N \\ 1 & \text{otherwise} \end{cases}.$$

DCT matrices are real and orthogonal. All DCTs are separable transforms; the multidimensional transform can be decomposed into a successive application of one-dimensional transforms (1-D) in the appropriate directions.

The DCT is a widely used frequency transform because it closely approximates the optimal KLT transform, while not suffering from the drawbacks of applying the KLT. This close approximation is based upon the asymptotic equivalence of the family of DCTs with respect to KLT for a first-order stationary Markov process, in terms of the transform size and the interelement correlation coefficient. Recall that the KLT is an optimal transform for data compression in a statistical sense because it decorrelates a signal in the transform domain, packs the most information in a few coefficients, and minimizes mean-square error between the reconstructed and original signal compared to

any other transform. However, KLT is constructed from the eigenvalues and the corresponding eigenvectors of a covariance matrix of the data to be transformed; it is signal-dependent, and there is no general algorithm for its fast computation. The DCT does not suffer from these drawbacks due to data-independent basis functions and several algorithms for fast implementation. The DCT provides a good trade-off between energy packing ability and computational complexity. The energy packing property of DCT is superior to that of any other unitary transform. This is important because these image transforms pack the most information into the fewest coefficients and yield the smallest reconstruction errors. DCT basis images are image independent as opposed to the optimal KLT which is data dependent. Another benefit of the DCT, when compared to the other image independent transforms, is that it has been implemented in a single integrated circuit (Rao and Yip 1990).

Research has displayed that among the family of DCTs, the performance of DCT-*II* is closest to the statistically optimal KLT based on a number of performance criteria. Such criteria include: energy packing efficiency, variance distribution, rate distortion, residual correlation, and possessing maximum reducible bits (Rao and Yip 1990). Furthermore, a characteristic of the DCT-*II* is superiority in bandwidth compression (redundancy reduction) of a wide range of signals and by existence of fast algorithms for its implementation. As this is the case, the DCT-*II* and its inversion, DCT-*III*, have been employed in the international image/video coding standards: JPEG for compression of still images, MPEG for compression of video including HDTV (High Definition Television), H.261 for compression telephony and teleconferencing, and H.263 for visual communication over telephone lines (Rao and Hwang 1996). Due to the superiority of

DCT-*II* in the family of DCTs, all subsequent discussion related to the DCT in this dissertation refers to the DCT-*II*.

1.3.3.1 One-dimensional DCT

. The one-dimensional DCT-*II* (1-D DCT) is a technique that converts a spatial domain waveform into its constituent frequency components as represented by a set of coefficients. The one-dimensional DCT-*III* is the process of reconstructing a set of spatial domain samples is called the Inverse Discrete Cosine Transform (1-D IDCT). The 1-D DCT has most often been used in applying the two-dimensional DCT (2-D DCT), by employing the row-column decomposition, which is also more suitable for hardware implementation. Figure 1.12 gives a visual representation of how the DCT works.

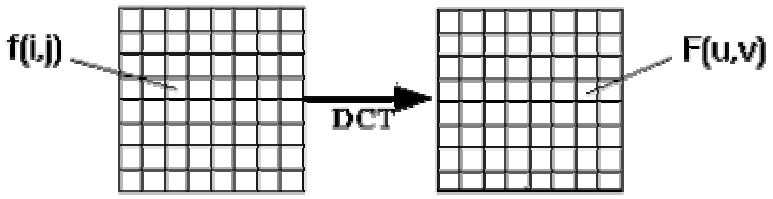


Figure 1.12: Generic 1-D DCT process.

The defining equation for the 1-D DCT is given by:

$$F(u) = \alpha(u)\sum_{x=0}^{N-1} f(x)\cdot\cos\left[\frac{\pi(2x+1)u}{2N}\right]$$

where $$\alpha(0) = \sqrt{\frac{1}{N}},$$

and $$\alpha(m) = \sqrt{\frac{2}{N}}.$$

Conversely, the defining function for the IDCT is given by:

$$f(u) = \sum_{x=0}^{N-1} \alpha(u)F(x)\cdot\cos\left[\frac{\pi(2x+1)u}{2N}\right]$$

where $\alpha(0) = \sqrt{\frac{1}{N}},$

and $\alpha(m) = \sqrt{\frac{2}{N}}.$

1.3.3.2 Two-dimensional DCT

The Discrete Cosine Transform is one of many transforms that takes its input and transforms it into a linear combination of weighted basis functions. These basis functions are commonly in the form of frequency components. The 2-D DCT is computed as a 1-D DCT applied twice, once in the *x* direction, and again in the *y* direction. The discrete cosine transform is of a *N* x *M* image *f(x, y)* is defined by:

$$F(u,v) = \frac{2}{\sqrt{MN}} \cdot \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} f(x,y) \cdot \cos\left[\frac{(2x+1)u\pi}{2N}\right] \cos\left[\frac{(2y+1)v\pi}{2M}\right]$$

with

$$\alpha(n) = \begin{cases} \frac{1}{\sqrt{2}}, & \text{for n} = 0 \\ 1, & \text{for n} > 0. \end{cases}$$

The inverse transform is defined by:

$$f(x,y) = \frac{2}{\sqrt{MN}} \cdot \sum_{u=0}^{N-1} \sum_{v=0}^{M-1} \alpha(u)\alpha(v)C(u,v) \cdot \cos\left[\frac{(2x+1)u\pi}{2N}\right] \cos\left[\frac{(2y+1)v\pi}{2M}\right].$$

In computing the 2-D DCT, factoring reduces the problem to applying a series of 1-D DCT computations. The two interchangeable steps in calculating the 2-D DCT are:

Step1: Apply 1-D DCT (vertically) to the columns.

Step 2: Apply 1-D DCT (horizontally) to result of Step 1.

Given in Figure 1.13 is a visual representation of how the 2-D DCT works.

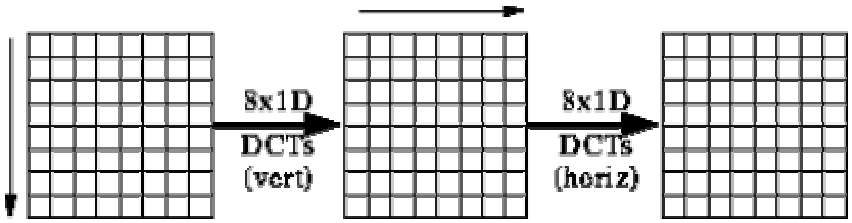


Figure 1.13: Generic 2-D DCT process.

In most compression schemes such as JPEG and MPEG, typically an 8x8 or 16x16 subimage, or block, of pixels (8x8 is optimum for trade-off between compression efficiency and computational complexity) is used in applying the 2-D DCT.

The DCT helps separate the image into parts (or spectral sub-bands) of differing importance (with respect to the image's visual quality). DCT transforms the input into a linear combination of weighted basis functions. These basis functions are the frequency components of the input data. For most images, much of the signal energy lies at low frequencies (corresponding to large DCT coefficient magnitudes); these are relocated to the upper-left corner of the DCT. Conversely, the lower-right values of the DCT array represent higher frequencies, and turn out to be smaller in magnitude, especially as u and v approach the subimage width and height, respectively.

1.3.3.3 JPEG Compression

One well known application of the DCT is the JPEG compression algorithm, established by the Joint Photographic Experts Group (Wallace 1990, Wallace 1991, Rao and Hwang 1996). The JPEG compression standard found motivation in the rapid growth of digital imaging applications, including desktop publishing, multimedia, teleconferencing, high-definition television (HDTV), and the increased need for effective and standardized image compression. JPEG compression is a lossy compression scheme for color and grayscale images. It works on full 24-bit color, and was designed to be used with photographic material and naturalistic artwork. Lossy compression is

compression in which some of the information from the original message sequence is lost. This means the original sequences cannot be regenerated from the compressed sequence. However, while some information is lost, that does not necessarily imply that the quality of the output is reduced. An example to support this claim is random noise. Random noise does not have very high information content, and when present in an image or a sound file, it is typically beneficial to drop it.

An overview of the JPEG compression algorithm is given in the following. The compression algorithm divides the image into 8x8 subimages, or blocks, of pixels. Each block produces a DCT array composed of 64 coefficients. These 64-component vectors from the separate blocks are compressed by a quantization step that places the coefficients into a discrete set of bins. Following coefficient computation, the bin numbers are transmitted. Upon reception, the cosine coefficient is approximated by the value at the middle of the bins.

DCT-based image compression schemes, such a JPEG, rely on two techniques to reduce the data required to represent the image. The first is *quantization* of the input image's DCT coefficients; the second is *entropy coding* of the quantized coefficients. Quantization is defined as the process of reducing the number of possible values of a quantity, thereby reducing the number of bits needed to represent it. Entropy coding is denoted as a technique by which the quantized data is represented as compactly as possible.

In the JPEG image compression standard, each DCT coefficient is quantized using a weight that depends on the frequencies for that coefficient. The coefficients in each 8x8 block are divided by a corresponding entry of an 8x8 quantization matrix, and the result is rounded to the nearest integer. In general, higher spatial frequencies are less

visible to the human eye than low frequencies. Therefore, the quantization factors are usually chosen to be larger for the higher frequencies. The *compression ratio* is a frequently used measure of how effectively an image has been compressed. This metric is given by the ratio of the size of the image file before and after compression. It is equal to the ratio of bit-rates, in bits/pixel, before and after compression.

1.3.3.4 MPEG Compression

Another well-known application of the DCT is MPEG (standard developed by the Moving Picture Experts Group (Rao and Hwang 1996) for video compression. The basic idea behind MPEG video compression is to remove spatial redundancy within a video frame and temporal redundancy between video frames. As in JPEG, DCT-based compression is used as a means to reduce spatial redundancy. In order to exploit temporal redundancy, *motion compensation prediction*[4] is applied. This method has shown good results because images in a video stream usually do not change much within small time intervals. The idea of motion compensation is to encode a video frame based on other video frames temporally close to it.

The MPEG compression method has two phases, encoding and decoding. Multimedia images are first encoded, then transmitted, and finally decoded back into a sequence of images by the MPEG player. The encoding process follows these steps:

1. The input data is transformed using a DCT.

2. The resulting DCT coefficients are quantized.

3. The quantized coefficients are packed efficiently using Huffman tables and run length coding.

[4] Motion-compensated prediction assumes that the current picture can be locally modeled as a translation of the pictures of some previous time.

Step 1 of the MPEG standard shows that at the heart of MPEG compression is the DCT, which has certain properties that simplify coding models and makes the coding efficient in terms of perceptual quality measures. For example, the advantage of applying the DCT can be observed when assuming a uniform 8x8 image block. In a uniform block (all pixels with the same value), only the first DC (Direct Current) coefficient of the DCT array is non-zero. As stated previously, the DCT is a vehicle for decomposing a block of data into a weighted sum of spatial frequencies (Intel 2002). During the decoding process, the MPEG player reverses these steps by unpacking the coefficients, dequantizing them, and applying a 2-D IDCT.

1.4 Summary

In this chapter, we have introduced the NoN paradigm, and its potential application in solving the problem of autonomic face recognition. The biologically plausible NoN approach for face recognition is a reasonable one due to the hierarchical manner in which spatial frequency information is used in the HVS. We discussed the widely used FFT, KLT, and DCT for use in transforming face images into a frequency representation, and due to the advantages over other transforms, the DCT was chosen as the transform to be used in this dissertation.

CHAPTER 2. FACE RECOGNITION AS PATTERN RECOGNITION

2.1 Introduction

In today's society, the application of autonomic image recognition schemes to face recognition from still images has become an important area of research due to the numerous applications for such technology, commercially and in areas of security and law enforcement. For robust applications of this nature, reliable algorithms for recognition, regardless of conditions at image capture such as lighting, facial orientation and expression, are required. Areas that show potential of face recognition systems include: entrance control in buildings, automatic teller machine (ATM) transactions, and criminal investigations.

While there are many areas of potential application for face recognition, other methods for recognition and authentication currently exist. However, typically, these existing methods are intrusive or require human action in the course of identification or authentication. An example of this is remembering and entering a password or personal identification number (PIN) code for authentication. Perhaps the greatest drawback to these methods of recognition is the susceptibility to imposter authentication due to lost or stolen passwords or PIN codes. Ignoring the special case of identical twins, it is reasonable to conjecture that in a robust face recognition system, authentication of an imposter would be far less probable when performing authentication due to the difficulty of feigning facial features.

There are at least two major categories of face recognition (Lawrence *et al*. 1996). The first category is finding a person within a large database of faces (e.g. in a F.B.I.

database). These systems typically return a list of the most likely people in the database. Often only one image is available per person. It is usually not necessary for recognition to be done in real-time. The second category involves identifying particular people in real-time (e.g. in a security monitoring system, location tracking system, etc.), or to allow access to a group of people and deny access to all others (e.g. access to a building, computer, etc.). Multiple images per person are often available for training and real-time recognition is required. A general implementation of a face recognition system is given in Figure 2.1.

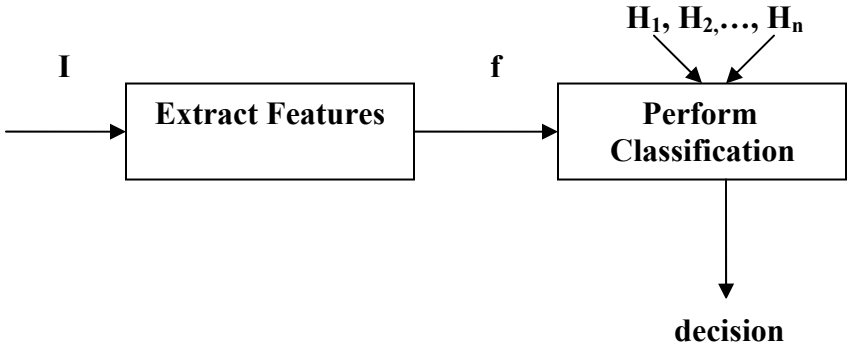


Figure 2.1: General face recognition system

The problem of face recognition can be stated as, given still or video images of a scene, identify one or more persons in the scene using a stored database of faces. In Figure 2.1, the test image is denoted as *I*, and upon completion of the feature extraction stage, the feature vector *f* is produced. Classification is performed by comparison of *f* to all hypothesis feature vectors $H_1, H_2, \ldots, H_n$ and a measure of similarity in the feature space is used to classify the test image as closest to one of the hypothesis vectors. System recognition performance is evaluated by the ratio of correct classification decisions over the total number of classification decisions made.

It is generally agreed that a well-defined and sufficiently constrained recognition problem will lead to a compact pattern representation and a simple decision making strategy. Learning from a set of examples is an important and desired attribute of most pattern recognition systems. The four best-known approaches for pattern recognition are template matching, syntactic matching, statistical classification, and neural networks.

Template matching is the earliest and simplest approach. A template or a prototype of the pattern to be recognized is matched against the pattern to be recognized. In the syntactic approach (e.g., Fu 1982, Pavlidis 1997) a pattern is viewed as being composed of simple sub-patterns which are themselves built from yet simpler sub-patterns, the simplest being the primitives. A complex pattern is then represented in terms of the interrelationships between these primitives. In the statistical approach, the patterns are described as random variables, from which class densities can be inferred. Classification is thus based on the statistical modeling of data. The neural network approach to pattern recognition is strongly related to the statistical methods, since they can be regarded as parametric models with their own learning scheme.

In this work, statistical pattern recognition will be performed, utilizing a neural network as a classifier. The recognition system, when given an input pattern, identifies the pattern as a member of a predefined class based on statistical analysis. Note that the recognition problem here is being posed as a classification task, where the classes are defined by the system designer (true for supervised classification). In order to describe a pattern, statistical techniques use quantitative values. The object is represented by a set of d features, or attributes, so it can be viewed as a d-dimensional feature vector. To make a classification, the space spanned by these feature vectors is subdivided using decision boundaries. Each part of the space then represents a class.

2.2 A General Probabilistic Framework

Given in this section is the general conceptual framework for face recognition in which the proposed system for autonomic face recognition was developed. We will examine two important subtasks of face recognition, namely, face recognition and face verification, as well as discuss important components of a face recognition system.

Face recognition, based on the desired manner in which recognition is to be performed, can be approached by handling the tasks of recognition or verification. Formally, given a database consisting of a set, ***F***, of *N* known people, *face classification* is the task of identifying the subject under the assumption that the subject is a member of ***F***. When considering the task of *face verification*, the subject's identity is supplied by some other means and must be confirmed. In the sections that follow, more specific information concerning these tasks are provided.

2.2.1 Subtasks of Face Recognition

2.2.1.1 Face Classification Tasks

The task of face classification task is an *N*-class classification problem, in which all *N* classes can be modeled. Here, we approach this task by collecting representative data for each of the *N* classes and applying a pattern classification technique. In general, the probability of misclassifying a face *x* is minimized by assigning it to the class C_k with the largest posterior probability $P(C_k \mid x)$, where

$$P(C_k \mid x) = \frac{p(x \mid C_k)P(C_k)}{p(x)}$$

$p(x)$ is the unconditional density, $p(x \mid C_k)$ is the class-conditional density and $P(C_k)$ is the prior probability for class C_k. We note that $p(x)$ is the same for every class, and it need not be evaluated in order to maximize posterior probability (Duda and Hart 1973).

Therefore, the classification task can be handled by modeling the class-conditional probability densities, $p(x \mid C_k)$, for each class.

Face classification is the recognition task addressed in this dissertation. The task of face classification is an *N*-class classification problem, where *N* represents the number of distinct individuals in the face image gallery. The collected representative data for each of the *N* classes is presented in the form of a set of face images used in training a classifier. A pattern classification technique is then applied, based on the backprogation learning algorithm. Here we minimize the probability of misclassifying a face *x,* thereby maximizing the probability of correctly classifying *x*, by assigning it to the class C_k with the largest posterior probability $P(C_k \mid x)$. This is accomplished by labeling *x* as a member of the class corresponding to maximum output posterior probability of the backpropagation algorithm.

2.2.1.2 Face Verification Tasks

Another task in face recognition is face verification, and this task can be viewed as a dual-class classification problem. The two classes, namely C_o and C_1, represent cases where the claimed identity is true and false, respectively. When considering this task, to maximize the posterior probability, *x* should be assigned to C_o if and only if the following condition holds true:

$$p(x \mid C_o) > \frac{p(x \mid C_1)P(C_1)}{P(C_o)}.$$

In this notation, the density $p(x \mid C_1)$ denotes the distribution of faces other than the claimed identity. Here we take the simplifying assumption that it is constant over the relevant region of space, falling to zero elsewhere. Under this assumption, the inequality above is equivalent to thresholding $p(x \mid C_o)$. However, it may be more

accurate to assume that the density $p(x|C_1)$ is smaller in regions of space where $p(x|C_o)$ is large. If $p(x|C_1)$ is chosen to be of the form $F[p(x|C_o)]$, where F is a *monotonically decreasing function*[5], then this assumption is also equivalent to thresholding $p(x|C_o)$. If this is the case, the threshold takes the form $G^{-1}\left[\frac{P(C_o)}{P(C_1)}\right]$, where $G(z) \equiv F(z)/z$. Since G is monotonic, G^{-1} is unique. If we only utilize data from class C_o, it is plausible to perform verification by thresholding $p(x|C_o)$. More accurate verification is attainable if negative data, i.e. data from class C_1, would need to be used in order to better estimate the decision boundaries. An iterative learning approach can be used in which incorrectly classified unknown faces are selected as negative data.

2.2.2 Functional Modules

A general face recognition algorithm can be divided into the following functional modules: a *feature extractor* that transforms the pixels of the facial image into a useful vector representation, and a *pattern recognizer* that searches the database to find the best match to the incoming face image. In considering the face segmentation problem, which will be discussed in detail later, the pattern recognizer categorizes the incoming feature vector into one of the two image classes: "face" images and "non-face images". However, in performing face recognition, the recognizer classifies the feature vector as belonging to a class of faces that is already registered in the database.

2.2.3 Feature Extraction

The input image of a face is typically modeled as a two dimensional array of numbers, i.e., pixel values. It can be written as $X = \{x_i, i \in S\}$ where $\boldsymbol{S}$ is a square lattice.

[5] Each member of a monotone decreasing sequence is less than or equal to the preceding member.

It may be more convenient to express X as a one dimensional column vector of concatenated row of pixels, $X = [x_1, x_2, ..., x_N]^T$ where ***N*** is the total number of pixels in the image. For images of size 92x112, which is the size of images used in this dissertation, N is as large as 10,304. Generally speaking, very high dimensional features are usually inefficient and also lack discriminating power. To handle the high dimensionality of the features, we must transform X into a *feature vector* $f(X) = [f_1(X), f_2(X), ..., f_M(X)]^T$ where $f_1(x)$, $f_2(x)$,...,$fm(x)$ are linear or non-linear functionals. Generally M is required to be much smaller than N in order to increase the efficiency of the new representation.

2.2.4 Pattern Recognition

Typically, the equations given in Section 2.2.3 can display random variations due to variants such as viewing angles, illumination, facial expression, and the like. As this is the case, it is generally better to model the vectors as random vectors. Assuming equal *a priori* probability, an incoming person is equally likely to be any person in the database. Therefore, by Bayes decision theory, the minimum recognition error rate can be achieved if the recognition is following the *maximum-likelihood* (ML) criterion. Specifically, if $Y = f(X)$ denotes the feature vector and there are K persons in the database, the identity of the incoming person is assigned by

$$k_o = \arg\min_{1 \le k \le K} \log p(Y \mid k)$$

where $p(Y \mid k)$ is the likelihood density of Y conditioning on its being the kth person. Here, we assume the variations in the facial feature vector are caused by zero-mean, additive white Gaussian noise, then the ML matching becomes the common *minimum distance* matching. This implies that the identity of the incoming person is k if the

Euclidean distance between the feature vector of the incoming person and the mean vector of the *k*th person is the smallest among all people in the database.

2.3 Groundwork and Literature Review

2.3.1 Introduction

There is much existing research on machine face recognition. In fact, research in this area has been conducted since the late 1960's. Due to important results of past literature, the resurgence of artificial intelligence studies in the 1980's, increased computational power available in computing systems of today, and potential applications for autonomic face recognition, face recognition research currently flourishes. In this section, an extensive review of the important groundwork in autonomic face recognition is performed. The literature review focuses on two aspects of machine vision in face processing; face segmentation, and face recognition, respectively. In Chapter 3, we move away from the legacy recognition systems discussed in this chapter and discuss recent research in machine face recognition.

2.3.2 Segmentation

We begin our discussion with groundwork in the area of face segmentation. Simply stated, face segmentation requires locating a face in an image (which can be noisy in nature). Typically, the task of finding a person's face in a picture requires little work for humans. This task, however, is non-trivial for machine vision systems. The significance of a robust solution to this problem is evident, in that segmentation holds an important key to future advances in human-to-human and human-to-machine communications. The segmentation of facial regions provides content-based representations of the images, and can be used for image processing in several areas, such as: encoding, pattern recognition and object tracking.

One of the earliest papers that reported the presence or absence of a face in an image was in Sakai *et al*. (1969). In this method, an edge map extracted from the input image is matched to a large oval template with possible variations in the position and size of the template. Following this, at positions where potential matches are reported, the head hypothesis is confirmed by inspecting the edges produced at expected positions. These expected positions correspond to prominent feature such as the eyes, mouth, etc.

Govindaraju *et al*. (1990) proposed locating the face in a cluttered image by applying a computational model. A deformable template is based on the outline of the head, working on the edge image. The template is composed of three segments that are obtained from the curvature discontinuities of the head outline. These three segments form the right side-line, the left side-line and the hairline of the head. In determining the presence of the head, all three of the segments are required to be in particular orientations. The center of the face is given by the location of the center of the three segments. The templates are allowed to translate, scale, and rotate according to certain spring-based models. A very small data set was used, approximately ten images, and though the author reports to have never missed a face, false alarms were reported.

Kelly (1970) introduced a top-down image analysis approach for automatically extracting the head and body outlines from an image in addition to the locations of eyes, nose, and mouth. In this approach, smoothed versions of original images were searched for edges that may form the outline of a head. Following this step, extracted edge locations were then projected back to the original image, and a fine search was locally performed for edges that form the head outline. The edges were connected, and once the head outline was obtained, the expected locations for eyes, nose and mouth were searched for locating these features.

A method for extracting the head area from an image based on the use of a hierarchical image scale and a template scale was proposed by Craw *et al*. (1987). A multi-resolution scheme was used, applying resolutions of 8x8, 16x16, 32x32, 64x64 and full scale 128x128 in scaling. At the lowest resolution a template was constructed of the head outline, using a line follower to connect the outline of the head. After determining the head outline, lower level features such as eyes, eyebrows, and lips were located, guided by the location of the head outline, using a similar method.

Working with both the intensity image of the face as well as the edge image found using the Canny's edge finder (Canny 1986), the face is segmented from a moderately cluttered background using an approach introduced in Sirohey (1993). In this approach, the preprocessing task included locating the intersection points of edges, and successively assigning labels to contiguous edge segments, thereby linking most likely similar edge segments at intersection points. The author approximated the human face using the ellipse as the analytical tool. Pairs of labeled edge segments L_i, L_j were fitted to the linearized equation of the ellipse (first equation given below). The author claims that this linearization is reasonable under the condition that the semi-major axis and/or the semi minor axis b of the ellipse are not 0. This assumption held true for all cases considered.

$$2x_i a_0 - y_i^2 a_1 + 2y_i a_2 - a_3 = x_i^2$$

where $a_0 = x_0, \qquad a_1 = \frac{a^2}{b^2}$

and $a_2 = \frac{a^2}{b^2} y_0, \qquad a_3 = x_0^2 + \frac{a^2}{b^2} - a^2.$

In the parameter set that resulted, x_0, y_0, a, and b are checked against the aspect ratio of the face. If this is satisfied, the set is included in the class of parameter sets for

final selection. The parameter set with the most segments is selected to represent the segmented face. A data set of 48 images was used, and a fairly good accuracy of above 80% was reported by the author.

Craw *et al*. (1992) described a hierarchical coarse-to-fine search system, motivated by automated indexing of police mug shots, to recognize and measure facial features. Forty feature points from a grayscale image were selected. The feature points were chosen according to Shepherd (1985), which was also used as a criterion of judgment. The template incorporated principles of polygonal random transformation, which was originally described in Grenander *et al*. (1991). The approximate location, scale and orientation of the head were obtained by iterative deformation of the whole template by random scaling, translation and rotation. A feasibility constraint was imposed so that these transformations did not lead to results that have no resemblance to the human head. An estimate of the location of the head was obtained, and then locality refinement was done by transforming individual vectors of the polygon. In a small data set of 50 images, the authors report successful segmentation over the entire set.

2.3.3 Recognition

2.3.3.1 Introduction

In recent times, an increasing focus on security measures in both the public and private sector has become evident. Within this environment of increased importance of security and organization, identification and authentication methods have developed into a key technological pioneer in various areas. Common identification and authentication applications include: entrance control in buildings, access control for computers, automatic teller machines (ATM) transactions, and use in the field of criminal investigation. It is reasonable to extend applications of this kind to others areas, such as

for use in intelligent personal computers. A possible application for this technology is making computers capable of interacting with the user on a level higher than mouse clicks and key strokes, rather on the basis of gesture and facial expression. Together with speech recognition software that is already marketed fairly inexpensively, highly intelligent man-machine interfaces could be built.

In this section of the dissertation, we will examine contributions in the area of face recognition. In performing face recognition, there are basic approaches, such as: classical, feature matching, neural network, and statistical approaches. A thorough review of work in face recognition using these approaches will follow.

2.3.3.2 Classical Approaches

Bledsoe reported one of the earliest works in computer recognition of faces in Bledsoe (1964). This method was not completely automated; in fact, a human operator located the feature points on the face and entered their positions into the computer. When fed the set of feature point distances of an unknown person, nearest neighbor or other classification rules were used for identifying the label of the test image. Due to manual feature extraction, the system handled variation head rotation, tilt, image quality, and contrast well.

Information theoretic arguments were employed in classifying human faces by Kaya *et al.* (1972). Here an upper bound on the entropy was calculated by reasoning to represent N different faces a total of $\log_2 N$ bits are required. The authors extracted prominent geometric features from the face image, and used a subset of these parameters to run statistical experiments on. They constructed a classifier based on the parameter vector and its estimate, i.e., if X is the parameter vector then the estimate Y is given as $Y = X + D$ where D is the distortion vector.

M. D. Kelly (1970) implemented a recognition system that required no human intervention, using the body and close up head images for recognition (this method used multi-resolution, or top-down image processing). The body and head were outlined, and ten measures were taken. Top down body measurements taken were height, width of the head, neck, shoulders, and hips. Facial distance measures included the width of the head, the distance between eyes, the top of head to the eyes, between eyes and nose, and from eyes to mouth. The nearest neighbor rule was used for identifying the class label of the test image.

Kanade (1977) utilized geometrical parameterization, a method of characterizing the face by distances and angles between points such as eye corners, mouth extremities, nostrils, and chin top. Face-feature points were located in two steps. The goal of the first step, a coarse-grain step, was to simplify the succeeding differential operation and feature-finding algorithms. The eyes, nose and mouth were approximately located, and then more accurate estimates were obtained by processing four smaller regions of the face, and scanning at higher resolution. The four regions were the left and right eye, nose, and mouth. Extracted were a set of sixteen facial parameters which represented ratios of distances, areas, and angles to compensate for the varying size of pictures. A small data set used in experimentation consisted of 40 images, using 20 of which for testing. Matching accuracies range from 45% to 75% correct, depending on the parameters used.

2.3.3.3 Feature Based Approaches

In Manjunath *et al*. (1992), proposed was a method for face recognition based upon feature points detected using the Gabor wavelet decomposition, and storing the results of the decomposition in a database. The goal was to facilitate recognition, as well

as to greatly reduce the storage requirement for the database. Here, 35 – 45 points per face image were generated and stored. The identification process utilized the information present in a topological graphic representation of the feature points. The method operated on controlled background images, such as passports and drivers license pictures. The method had a 94% recognition rate on a small database, but did display a dependency on the illumination direction in the image.

2.3.3.4 Neural Network Approaches

One of the most widely used unsupervised neural network architectures is that of the Kohonen neural network. The Kohonen associative map (Kohonen 1988) was one of the earliest neural network algorithms used in the area of face recognition. Using this method on a small data set, accurate recall was reported even when portions of the input image was noisy or missing, as demonstrated by Psaltis's group (Abu-Mostafa and Psaltis 1988), using optical hardware.

In Seibert and Waxman (1991), the authors proposed a system for recognizing faces from their parts using a neural network. This modular based system combined 2-D views from different vantage points, arranging the features so that prominent features such as eyes and nose played the a dominant role in the 2-D views. There were several steps involved in processing a face for recognition. Namely, segmentation of a face region using inter-frame change detection techniques, extraction of features such as eyes, mouth, etc., using symmetry detection, grouping and log-polar mapping of features and their attributes such as centroids, encoding of feature arrangements, clustering of feature vectors into view categories using ART 2, and integration of accumulated evidence using an aspect network. Seibert and Waxman furthered their work in (Seibert and Waxman

1993), attempting to model the role of caricatures and in human face recognition to enhance the capabilities of their recognition system.

Stonham *et al*. (1984) introduced a recognition device named WISARD. In the purest form, the system was single layer adaptive NN implementation for face recognition, expression analysis and face verification. As a fairly computationally expensive system, the process required typically 200-400 presentations for training each classifier. These training patterns included translation and variation in facial expressions, in hopes of improving recognition. Sixteen classifiers were used for the dataset constructed using 16 persons.

A two-stage neural network classifier is used in Golomb and Sejnowski (1991), for gender classification. The first stage is an image compression NN whose hidden nodes serve as inputs to the second NN that performs gender classification. The networks consisted of fully connected, three-layer networks with two biases and are trained by a standard back-propagation algorithm. The testing and training set was kept as simple as possible, having no facial hair, jewelry or makeup in a face image. The compression network stage extracted features from the input image; the second network actually performed the gender classification. Later work with this system was extended to classifying facial expressions as well.

Brunelli and Poggio (1992) proposed the use of two HyperBF networks (Poggio and Girosi 1990) for gender classification based on face images. The authors used a vector of 16 numerical attributes (e.g., eyebrow thickness, widths of nose and mouth, etc.). In this work, two HyperBF networks were trained, one corresponding to the male and female gender. The authors normalized the input images scale and rotation by using the positions of the eyes which were automatically detected. The output of both

HyperBF networks were compared, and the network with the largest output magnitude was declared the winner, more specifically, the image was labeled male or female based on which network won. The database consisted of 21 males and 21 females. The leave-one-out strategy (Fukunaga 1989) was employed for classification. A recognition accuracy of 92.5% was reported when the feature vector from the training set was used as the test vector and 87.5% when test were run on images not in the training set.

A system for face recognition in still monochromatic images was proposed by Huang *et al*., based on a neural network like structure called a Cresceptron (Yang and Huang 1993). The Cresceptron architecture was essentially a multi-resolution pyramid structure network the employs automatic, incremental learning. Before applying Cresceptron learning, the face images were first operated on by a rule-based algorithm to locate faces in the image (Yang and Huang 1993). In a small-scale experiment involving 50 persons, the Cresceptron performed well.

In Brunelli and Poggio (1992), the use of HyperBF networks for face recognition was examined. In this work, the images were first transformed using 2-D affine transforms in order to remove variations due to changing viewpoint. Detected positions of the eyes and mouth in the face image were used to obtain the transformation features as well as the positions of these features. Following this, the effects of illumination in the image was reduced by applying a directional derivate operator to the transformed image. The result was multiplied by a Gaussian function and integrated over the receptive field. The goal of this step was to achieve dimensionality reduction. In experimentation, the MIT Media Lab database of 27 images was used, with the images of 17 persons being used for training, while the rest were used as testing samples. The authors reported an

average accuracy of 79% on the database. The authors claimed higher accuracy rates when applying the outputs of 16 HyperBF networks into another HyperBF.

A connectionist model for face expression recognition was proposed in Rahardja *et al*. (1991). Here, the model uses the pyramid structure to represent image data, where each level of the pyramid was represented by a network. The network consisted of a single input, hidden, and output layer. The learning algorithm used was a variation of the back propagation learning algorithm. For experimentation, an interesting data set for training was comprised, using hand drawn faces with six different expressions: happiness, surprise, sadness, anger, fear, and normal. The faces were drawn in such a way to be as dissimilar as possible from others. The testing set consisted of only six training faces. The network was able to successfully recognize the members of the training set when tested on them. However, the network poorly recognized the training images when applied as tests when the images were slightly modified by blurring or distorting the facial expressions.

Systems based on dynamic link architectures (DLAs) are proposed in Buhmann *et al*. (1990) and Pentland *et al*. (1990). The goal of employing a DLA based system is to solve the conceptual problem of representing syntactical relationships in neural networks. In DLA based systems, *synaptic plasticity*[6] is used, and the networks are able to instantly form sets of neurons grouped into structured graphs, while retaining the benefits of neural systems. Both Buhmann *et al*. (1990) and Pentland *et al*. (1990) use Gabor wavelets for the feature extraction, keeping information concerning frequency, position, and orientation.

[6] Synaptic plasticity models how synaptic connections can be altered by prior neuronal activity.

2.3.3.5 Statistical Approaches

A statistical based approach based on the standardized KL coefficients was proposed in Akamatsu *et al*. (1991). Two experiments were performed using this method. The test set consisted of five samples from 20 individuals. In the first experiment, the training and testing samples were acquired under near uniform conditions (lighting, orientation, etc.). The second experiment checked for feature robustness when there is a variation caused by an error in the positioning of the target window. The authors noted that this error occurred during image acquisition due to changing conditions. In the second experiment, the test images were created by shifting the reference points in various directions by one pixel. The variances for 4 and 8 pixels were tested. The experiments had accuracy rates of 85% and 91%, respectively.

Eigenpictures (also referred to as "eigenfaces" in the context of face recognition) were proposed by Turk and Pentland (1991) for identification. The authors suggested that given the eigenfaces, every face in the database can be represented as a vector of weights; the weights are obtained by projecting the image into eigenface components by a simple inner product operation. In this method, when a new test image whose identification is required was given, the new image was also represented by its vector weights. Location of an image in the database, whose weights were the closest (in Euclidean distance) to the weights of the test image, was the means by which identification of a test image was handled. The authors reported good results on an their large database of face images, which consisted of 2500 face images of 16 subjects, digitized at all combinations of three head orientations, three head sizes and three lighting conditions. Experiments were run to determine the ability to handle variation in lighting, size, head orientation, and the differences between the training and test conditions. The

authors reported 96% correct classification over lighting variations, 85% over orientation variations and 64% over size variations. Turk and Pentland also applied their method to real time recognition of a moving face image in a video sequence. Here, spatiotemporal filtering was performed, followed by a non-linear operation used to identify a moving person.

As an application of their recognition scheme, the authors devised a system for interactive search through a face database. The goal of the system was to return face images of certain types that were similar to the user descriptive query. For example, if the user input was "White male of age 35 years or younger", images that satisfied this query were returned in groups of 21. If the user selected one of the returned images, the system presented faces from the database that looked similar to the chosen face in the order of decreasing similarity. In a test involving 200 selected images, approximately 95% recognition accuracy was obtained. A second experiment was run to evaluate the recognition accuracy as a function of race. Here, images of white, black and Asian adult males were tested. For white and black male accuracies of 90% and 95% were reported, respectively, while a slightly lesser accuracy 80% was reported for Asian males.

In Chen *et al*. (1987), the Singular Value Decomposition (SVD) was proposed for face recognition. In this method, the SV feature vector is compressed into a low dimensional space by means of an optimal discriminant transform based on Fisher's criterion. The optimal discriminant transform compresses the high-dimensional SV feature space to a new *r*-dimensional feature space. The new secondary features are algebraically independent and informational redundancy is reduced. Goshtasby's shape matrices were used to represent the images, which are rotation, translation, and scale invariant of the facial images. These shape matrices were obtained by polar quantization

of the shape (Goshtasby 1985). This approach was tested on 64 facial images of eight people, where a separate class corresponded to each person, in performing recognition. Three photographs from each class were used to provide a training set of 24 SV feature vectors. The SV feature vectors were treated with the optimal discriminant transform to obtain new feature vectors for the 24 training samples. The remaining 40 facial images were used as the test set, five from each person. The results on this small data set were good, at 100% recognition.

Cheng *et al* (1992) also performed research using SVD methods. Particularly, an algebraic method for face recognition using SVD and threshold value was developed. Projective analysis was used with the training set of images serving as the projection space. A training set in their experiments consists of three instances of face images of the same person. In this approach, eigenvalues and eigenvectors were determined for the average image. Following this, thresholding of the eigenvalues was performed in order to disregard the values close to zero. A test image is then projected onto the space spanned by the eigenvectors. The authors use the *Frobenius norm*[7] as criterion to perform classification. When working with a database of 64 face images of eight different persons, the authors reported 100% accuracy. In comprising the data set, each person contributed eight images. Three images from each person were used to determine the feature vector for the face image in question.

In Nakamura *et al*. (1991), the use of *isodensity lines*[8] for face recognition was proposed. The author claimed that these lines provide a relief image of the face, while they are not directly related to the 3D structure of a face. The Sobel operator and

[7] The Frobenius norm of a matrix **A** equals sqrt(sum(**ABS**(**A**)^2)) and is written $||\mathbf{A}||_F$. It is always a real number.

[8] Curves of constant gray level

necessary post-processing steps were utilized in obtaining the boundary of the face region. To trace contour lines on isodensity levels, the gray level histogram was then used. Template matching was used as a final step for face recognition. A small data set of ten pairs of face images, with three pairs of pictures of men with thin beards, and two pairs of pictures of women was used. On this small data set, the author reported 100% recognition accuracy.

2.4 Summary

Given in this chapter was the general framework in which the proposed system for autonomic face recognition was developed, in addition to a review of groundwork in the area of machine vision of faces. In the literature review, we have discussed the two areas, namely, segmentation and recognition. The first area discussed was face segmentation, which is the process of finding a face in an input image. In images with fairly uniform backgrounds (this is the case for many photos used for personal identification), where only the face is present; many legacy systems performed well, based on accuracies claimed by the authors. Literature has shown that error rates in segmentation increase when the face images were in noisy or cluttered scenes, however. The second application reviewed was face recognition, which is the focus of this dissertation. The general idea of face recognition is to extract the relevant information in a face image, encode it as efficiently as possible, and compare one face encoding with a database of similarly encoded images.

Several legacy methods for segmentation and recognition were discussed. In many of the previous works, the authors generated face image databases for training and testing of their proposed systems. The problem with this was many of these databases were very small (typically less than 50 images) and were not publicly available.

Determining the merit of a given system, or direct comparison between systems was extremely difficult, as no benchmarking was possible with varying data. Many of the legacy systems published fairly high recognition rates, on very small image galleries (database of face images), but when others attempted to recreate the results on larger databases, the recognition rates greatly decreased.

CHAPTER 3. RECENT RESEARCH IN FACE RECOGNITION

3.1 Introduction

Modern algorithms for face recognition have shown promise, reaching an identification rate of greater than 90% for larger databases with well-controlled pose and illumination conditions. However, even with such notable levels of accuracy, statistically, face recognition as a means for identification and authentication is not comparable to methods using keys, badges or passwords, nor can it bear direct comparison with the recognition abilities of a human concierge (Fromherz *et al*. 1997). Still, face recognition as a means of identification and authentication is becoming more plausible with continual research contributions in the area.

3.2 AT&T Cambridge Laboratories Face Database

The AT&T Cambridge Laboratories face database (formerly the ORL face database), was built at the Olivetti Research Laboratory in Cambridge, UK and is available free of charge from http://www.uk.research.att.com/facedatabase.html. The database consists of 400 different images, 10 for each of 40 distinct subjects. There are 4 female and 36 male subjects. For some subjects, the images were taken at different times, varying the lighting, facial expression (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position with tolerance for limited side movement and limited tilt up to about 20 degrees. There is some variation in scale of up to about 10%. The size of each image is 92x112 pixels, with 256 grey levels per pixel.

Sample thumbnails of images in the AT&T Cambridge Laboratories face database can be seen in Figure 3.1.

Figure 3.1: A sample of images in the AT&T Cambridge Laboratories face database

3.3 Comparison of Results

Given in Table 3.1 is a comparison of face recognition error rates achieved when performing testing on the AT&T Cambridge Laboratories face database in recent literature. The works cited in this section are recent, citing works no earlier than 1994, having most results provided from work done in the 2000s. In all works cited, when performing testing on the AT&T Cambridge Laboratories face database, the first five images of an individual were chosen for training, and the other five for testing (i.e., a total of 200 testing images). The error percentages in Table 3.1 represent the author's best mean reported performance. In Table 3.1, the references are ranked by error rate. This allows the reader to easily compare the results of the methods over the database. In this dissertation, the AT&T Cambridge Laboratories face database is also used for testing. Details of the results obtained when utilizing the proposed approach will be given later in Chapter 5.

Table 3.1: Comparison of error rates in current literature when using the AT&T Cambridge Laboratories face database.

System	Error %
Gaussian Weighting (Hjelms 2000)	15.0
Hidden Markov Models (Nefian and Hayes 1998)	14.0
Eigenface (PCA) (Samaria 1994)	10.0
Linear Discriminant Analysis (Yu and Yang 2001)	9.2
Low frequency DCT with subimages (Pan and Bolouri 1999)	7.35
Low frequency DCT w/o subimages (Pan and Bolouri 2001)	5.85
Pseudo-2D Hidden Markov Models (Samaria and Young 1994)	5.0
Probabilistic decision-based neural network (Lin *et al*. 1997)	4.0
Convolutional neural network (Lawrence *et al*.1997)	3.8
Linear support vector machines (Guo *et al*. 2001)	3.0
Kernel principle component analysis (Kim *et al*. 2002)	2.5
One Spike Neural Network (Delorme and Thorpe 2001)	2.5
Uncorrelated Discriminant Transform (Zhong *et al*. 2001)	2.5

3.4 Discussion

3.4.1 Methods

In Hjelms (2000), the goal is to perform face recognition using local features of a face image, rather than holistic features, using Gabor coefficients. An input face image is filtered with a set of Gabor filters. The filtered image is then multiplied with a 2-D Gaussian to focus on the center of the face, avoiding features at the face contour. The Gabor filtered and Gaussian weighted image is then searched for peaks, which the author retains as feature points. These peaks are extracted as features, and the location and class label are also stored. In this approach, Gabor wavelets extract facial features including an entire neighborhood, similar to the receptive fields of the human visual system, thus representing the face as a collection of feature points (Fromhertz *et al.* 1997).

Nefian and Hayes (1998) and Samaria and Young (1994) use Hidden Markov Models (HMM) for face recognition. HMM are a set of statistical models used to characterize the statistical properties of a signal (Rabiner and Huang 1993). HMM consist of two interrelated processes (Nefian and Hayes 1998): (1) an underlying, unobservable Markov chain with a finite number of state, a state transition probability matrix and an initial state probability distribution and (2) a set of probability density functions associated with each state. HMM generally work on sequences 1-D feature vectors, while an image usually is represented by a simple 2-D matrix. To handle this, a sliding window is applied to the image, covering the entire width of the image, which is moved from the top to the bottom of the image. The brightness values of the windows are passed to the HMM process as 1D-feature vectors. Successive windows overlap to avoid cutting off significant facial features and to bring the missing context information into the sequence of feature vectors. The human face can be divided in horizontal regions like forehead, eyes, nose, etc. that are recognizable even when observed in isolation. Thus the face is modeled as a linear left-right HMM model of five states, namely forehead, eyes, nose, mouth, and chin.

The complete recognition system incorporates a preprocessing step followed by four processing steps (Fromhertz, *et al.* 1997): (1) The feature vectors of a test set of images is used to build a code book; (2) Based on this code book all feature vectors are quantized; (3) The HMM model is trained for every person in the database; (4) Recognition: A test image, not used for training, passes the preprocessing step and step 1 and 2, before the probability of producing this image is computed for every person, i.e. every model, in the database. This procedure produces a ranking of the potential individuals in descending order of the probability of each model.

In Samaria (1994) and Kim *et al*. (2002), face recognition based on PCA, or Karhunen-Loeve expansion, is performed. Perhaps the best known unsupervised feature extractor, PCA computes the d largest eigenvectors from a $D \times D$ covariance matrix of the n D-dimensional patterns. Since PCA uses the most expressive features (eigenvectors with largest eigenvalues), it effectively approximates the data by a linear subspace using the mean squared error criterion (De Baker 2002). PCA-based approaches typically include two phases (Yang *et al*. 2000): training and classification. In the training phase, an eigenspace is established from the training samples using the principal components analysis method. The training face images are then mapped onto the eigenspace. In the classification phase, the input face image is projected to the same eigenspace and classified by an appropriate method. PCA-based schemes have shown promise when beards, baldness, or non-uniform backgrounds are present in the imagery, since it treats the entire image, face and background, as a holistic pattern. To improve on classification results obtained with PCA, Kim *et al*. (2002) recently proposed a nonlinear extension. The basic idea is to first map the input space into a feature space via a nonlinear mapping and then compute the principal components in that feature space.

An algorithm for face recognition based on Linear Discriminant Analysis (LDA) is proposed by Yu and Yang (2001). There are typically two steps in a LDA-based face recognition system: (1) Perform PCA to project the n-dimensional image space onto a lower dimensional subspace; (2) Perform discriminant projection using LDA. The first step of PCA is necessary because the standard LDA algorithm has difficulty processing high dimensional image data. To handle this, PCA is often used for projecting an image into a lower dimensional space (referred to as the face space), and then LDA is performed to maximize the discriminatory power. Here, PCA plays a role of dimensionality

reduction and forms a PCA subspace (Yang *et al*. 2000). As an optimization, Yu and Yang (2001) propose a direct LDA algorithm for face recognition that accepts high-dimensional data (face images), but does not require PCA-based dimensionality reduction.

A probabilistic decision-based neural network is proposed by Lin *et al*. (1997). In this work, the authors introduced an efficient probabilistic decision based NN (PDBNN) for face detection and recognition. Feature vectors were generated using intensity and edge values obtained from a facial region of the down sampled images in the training set of face images. The facial region contains the eyes and nose, but excludes the hair and mouth. Two PDBNNs were trained with these features vectors and used one for face detection and the other for face recognition.

Lawrence *et al*. (1997) proposes a convolution neural network for face recognition. Here, feature vectors are generated in one of two ways. The first method is similar to the HMM local image sampling process of stepping a window over the face image, generating 1-D feature vectors at each location. A second method creates a representation of the local sample by forming a vector out of: (a) the intensity of the center pixel, and (b) the difference in intensity between the center pixel and all other pixel within a square window (Lawrence *et al*. 1997). The resulting representation becomes partially invariant to variations in intensity of the complete sample. The degree of invariance is modified by adjusting the size of the window. Feature vectors based on the above methods are then fed to the convolution NN for classification.

The uses of Support Vector Machine (SVM) classification is used by Guo *et al*. (2001). SVMs belong to the class of maximum margin classifiers. They perform pattern recognition between two classes by finding a decision surface that has maximum distance

to the closest points in the training set which are termed support vectors. In solving *n*-class classification, there are two basic strategies (Heisele *et al*. 2001): (1) In the one-versus-all approach *n* SVMs are trained. Each of the SVMs separates a single class from all remaining classes; (2) In the pairwise approach $n(n-1)/2$ machines are trained. Each SVM separates a pair of classes. The pairwise classifiers are arranged in trees, where each tree node represents an SVM. The latter strategy is used by Guo *et al*. (2001), specifically; a binary tree structure is used. For a coming test face, it is compared with each two pairs, and the winner will be tested in an upper level until the top of the tree. By bottom-up comparison of each pair, the unique class number will finally appear on the top of the tree.

Delorme and Thorpe (2001) approach face recognition from a biological viewpoint, using a one spike neural network. The goal of the one spike neural network is to model the manner in which visual information is propagated through the human vision system. The premise is as follows: visual information, in the form of signal spikes, has to travel through many layers in the visual system. Vision research implies that the neurons at most processing stages will rarely be able to fire more than one spike before the next stage has to respond. As this is the case, the author proposes the use of a network that uses only one spike per neuron to process faces. The author ran simulations using SpikeNet (Delorme *et al*. 1999), a software package designed for modeling networks of this kind.

A method to directly extract discriminant features for face images by using the uncorrelated discriminant transformation and PCA is proposed by Zhong *et al*. (2001). This approach is very similar to the LDA approach. As stated previously, in LDA-based approaches, PCA is applied as an initial step for dimensionality reduction. Zhong *et al*.

(2001) proposes a transform that can be applied under certain conditions, the uncorrelated discriminant transformation (Jin *et al*. 1999), to eliminate this step. Specifically, suppose there are L known face classes. If the face images have a resolution m x n then the dimensionality of the original space is $N = mn$. The author takes K to represent the number of training samples. If the resolution m x n is so low that $K > N + L$, then the uncorrelated discriminant transformation can be directly used to extract features of face images, otherwise PCA must first be applied (Zhong *et al*. 2001).

The use of DCT coefficients is proposed by Pan and Bolouri (1999), (2001). In these works, the DCT is applied to a face image, and feature vectors of variable size, based on varying numbers of retained low frequency coefficients, are used in classification. The retained coefficients are fed to a multilayer perceptron classifier. An in-depth discussion of this approach is deferred to Chapter 5, when a comparison between this approach and the method proposed in this dissertation will be performed. It will be displayed that the method for face recognition proposed in this dissertation improves upon the results obtained in Pan and Bolouri (1999), (2001).

3.4.2 Biological Motivation

A subset of the algorithms listed in Section 3.4.1 are biologically motivated (e.g., Samaria 1994, Hjelms 2000, and Delorme and Thorpe 2001). In this dissertation, we also propose a biologically motivated method for face recognition, based on the NoN model, which performs recognition by using spatial frequency information of a face image in a hierarchical fashion, which is analogous to the way faces are processed in the HVS. This dissertation will show that recognition rates, in comparison to those achieved by biologically motivated schemes or otherwise, are comparable to or better than, most approaches.

3.5 Summary

In this chapter, a review of recent research in face recognition was performed. The AT&T Cambridge Laboratories face database was introduced as a benchmarking data set, and results of varying systems were given, when performing testing using this database. As given in Table 3.1, the methods based on Kernel PCA, the One Spike Neural Network, and the Uncorrelated Discriminant Transform, performed best, with an error rate of 2.5%. Delorme and Thorpe (2001) report very good results based on the combination of a dedicated network for face recognition and robust features based on Gabor coefficients. The other two best results over the database, Zhong *et al*. (2001) and Kim *et al*. (2002), are PCA based schemes that employ the optimal, yet computationally costly, Karhunen-Loeve expansion. It is plausible that classification accuracy obtained by methods based on KL expansion will obtain good classification results, due to the optimal energy compaction property of the expansion.

CHAPTER 4. FEATURE SPACES

4.1 Feature Vectors and Related Topics

4.1.1 Introduction

Given a pattern, its recognition may consist of one of the following two tasks: (1) supervised classification, in which the input pattern is identified as a member of a predefined class, (2). unsupervised classification, in which the pattern is assigned to a hitherto unknown class (De Baker 2002). In this context, the recognition problem is given as a classification or categorization task, where the classes are either defined by the system designer (supervised classification) or are learned based on the similarity of patterns (unsupervised classification). This dissertation emphasizes on supervised pattern recognition methods.

In describing a pattern, neural networks and statistical techniques use quantitative values. Hence, the object is represented by a set of d features, or attributes, so it can be viewed as a d-dimensional feature vector. In making a classification decision, the space spanned by these feature vectors, is subdivided using decision boundaries. A class is represented by each part of the space. To establish these decision boundaries, concepts from statistical decision theory are utilized. In Figure 4.1 (De Baker 2002), a schematic representation of a recognition process is shown. In the diagram, a preprocessing module can be applied to segment the pattern of interest from the background, remove noise, normalize the pattern, and any other useful form of preprocessing, which will improve the compacted representation of the pattern. A recognition system can be operated in two modes: training and classification. In the training mode the feature extraction/selection

module finds the appropriate features for representing the input patterns. The classifier is trained to partition the feature space, and feedback path allows the designer to optimize the pre-processing and feature extraction/selection methods. In the classification mode, the trained classifier assigns the input pattern to one of the pattern classes, based on the measured features. The dataset used during construction of the classifier, the training set, is different from the one used for performance estimation, the testing set.

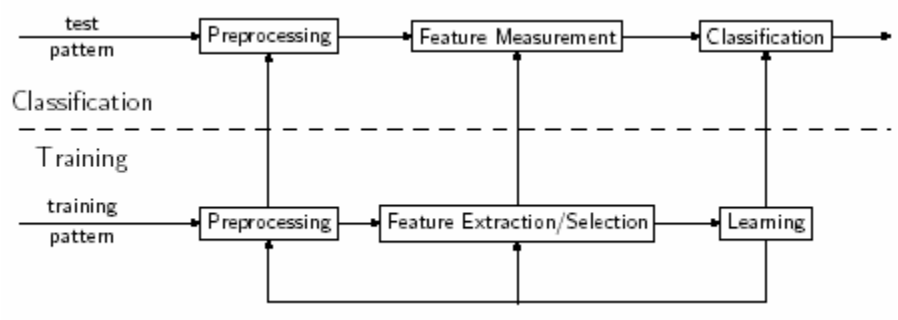


Figure 4.1: A general recognition process

The decision making process in pattern recognition under consideration can be summarized as follows (De Baker 2002): A given pattern is to be assigned to one of L categories $C_1, C_2, ..., C_L$ based on a vector of D feature values $x = (x_1, x_2, ..., x_D)$. Statistically, the features are assumed to have a probability density or mass function conditioned on the pattern class. Thus, a pattern vector x belonging to class C_i is viewed as an observation drawn randomly from the class-conditional probability function $f(x|C_i)$. To define the "optimal" decision rule, the Bayes decision rule is used.

4.1.2 Features

Features are representative characteristics extracted from annotated objects to be classified in a recognition system. In most recognition systems, these features are group in a 1-D array, forming a feature vector. Feature vectors characterizing visual objects can be as simple as raw pixel values, a histogram or distribution of intensities, or perhaps

intensity profiles along relevant axis or their gradients. More advanced features can include components from wavelet and FFT transforms. In a supervised learning system, once the classifier is trained with examples, it is capable of generalization and can consequently classify situations never seen before by associating them to similar learned situations. Given in Figure 4.2 is example of feature extraction from a face image and the decision space mapping for a general face recognition system.

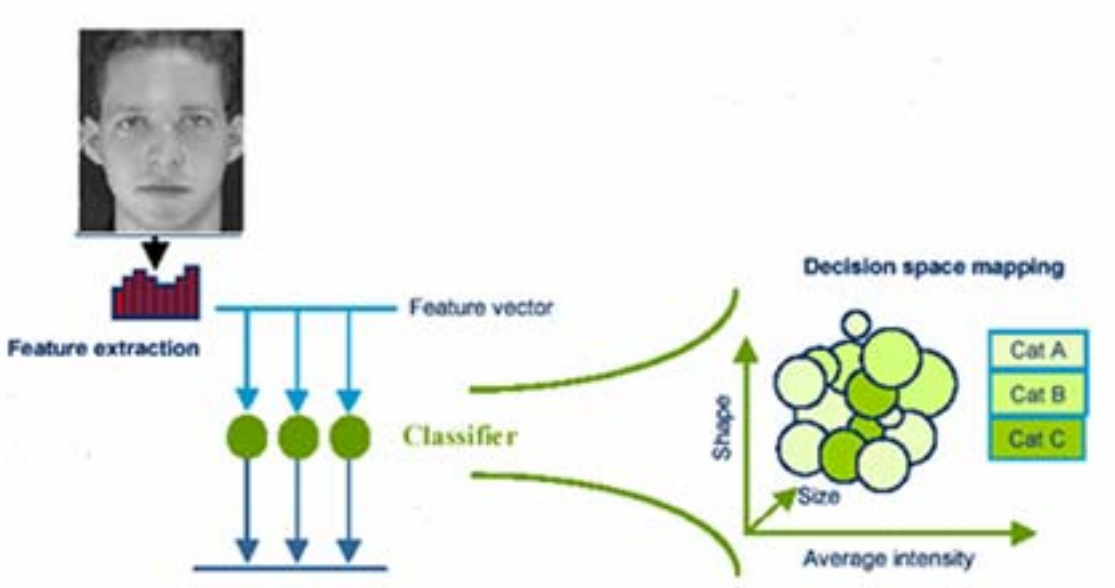


Figure 4.2: A general decision space mapping.

In the event that several features are used to accurately identify a population of objects, each one of them contributes to the generation of a recognition engine or sub-engine. Their diagnostics can then be weighted and consolidated for final decision.

4.1.3 Feature Selection and Extraction

Given a set of example feature vectors with their corresponding class labels, dimensionality reduction can be achieved in essentially two different ways: feature selection and feature extraction. The term *feature selection* refers to algorithms that select the best subset of the input feature set. Feature selection will identify and neglect

those variables that do not contribute to the classification task. The task is to seek d features out of the available D measurements. For feature selection, the resulting classification error is used as a measure for good selection. However, this makes the feature selection procedure dependent on the specific classifier that is used. More formally, the feature selection problem involves, first, examining all $\binom{D}{d}$ possible subsets of size d. Secondly, selection requires taking the subset with the smallest classification error. Note that the number of possible subsets grows combinatorially. This makes the search impractical for even moderate values of D and d. A feature selection method has been proposed by Narendra and Fukanaga (1977), based on the branch and bound algorithm, which does avoid this exhaustive search, however. The key to this algorithm is the monotonicity property of the criterion function $J(\cdot)$, which can formally be stated as: given two features subsets X_1 and X_2, if $X_1 \subset X_2$ then $J(X_1) < J(X_2)$. Simply put, the main idea of the work states that the performance of a feature subset should improve whenever a feature is added to it. The most commonly used classification functions do not satisfy this monotonicity property.

Feature extraction is performed by a technique called discriminant analysis. Methods that create new features based on a transformation or combination of the original feature set are called feature extraction algorithms. Here, the category information associated with each pattern is used for linearly extracting the most discriminatory features. The interclass separation is emphasized by finding the eigenvectors of a general separability measure, like the Fisher criterion $S_w^{-1}S_b$ (the product of the inverse of the within-class scatter matrix, S_w, and the between-class scatter

matrix, S_b) (Fukunaga 1990). The eigenvectors with the largest eigenvalues will give the best separating features.

The best known unsupervised feature extractor is PCA, which computes the d largest eigenvectors from $D \times D$ covariance matrix of the n D-dimensional patterns. Since PCA uses the most expressive features (eigenvectors with largest eigenvalues), it effectively approximates the data by a linear subspace using the mean squared error criterion. Other methods, like projection pursuit (Friedman 1987) and independent component analysis (ICA) (e.g., Comon 1994, Bell and Sejnowski 1995, Cardoso 1998, Lee 1998, Hyvarinen *et al*. 2001) do not rely on the second-order properties of the data and are more appropriate for non-Gaussian distributions.

4.1.4 Dimension Minimization

In many classification problems, high-dimensional data are involved, because large feature vectors are generated to be able to describe complex objects, and to distinguish between them. While this is the case, the number of actual available data points is limited in many practical situations. For a classifier, the estimation of the class probability distributions in these sparsely sampled high-dimensional data spaces is non-trivial, and generally affects the correctness of the obtained classification results. To avoid these problems, the dimension of the feature space is reduced. This can be done in several ways. The easiest way is to select a limited set of features out of the total set (Devijver and Kittler 1982). The classification performance serves as a measure for selecting the features.

Feature extraction is a widely used method for selecting a limited set of features out of the total set, in approaching dimension minimization. Here, features are extracted as functions (linear or non-linear) of the original set of features. Unsupervised linear

feature extraction techniques generally are PCA-based schemes, which rotate the original feature space, before projecting the feature vectors onto a limited number of axes. Supervised feature extraction techniques usually relate to the discriminant analysis technique (Fukunaga 1990), which uses the within and between-class scatter matrices.

Performing dimension minimization, in order to keep the dimensionality of an input pattern representation as small as possible, is advantageous for reasons such as minimizing computational cost and improving classification accuracy. A low-order yet representative feature set simplifies the pattern representation, the classifiers that are built on the selected representation, and necessary computations. As this is the case, the resulting classifier will be faster and will use less memory. A small number of features can alleviate the *curse of dimensionality*[9] when the number of training samples is limited. However, a reduction in the number of features may directly lead to a loss in the discrimination power of a classifier and thereby yield lower accuracy in recognition.

An important issue in dimensionality reduction of feature vectors is the choice of a criterion function. A commonly used criterion is the classification error of a feature subset. This is not a simple task to accomplish because the classification error itself cannot be reliably estimated when the ratio of sample size to the number of features is small. In addition to the choice of a criterion function, a second issue is determining the appropriate dimensionality of the reduced feature space.

In performing classification based on extracted features, a set of decision boundaries are built, using only a finite set of example objects. This limited set constrains the learning process, in the sense that it is important to limit the number of features fed into the classifier. The classifier is trained using the available training samples, and

[9] The Curse of dimensionality (Bellman, (1961)) refers to the exponential growth of a hypervolume as a function of dimensionality.

performance of the classifier depends on the sample size, number of extracted features, and the selected classifier. Classifier design involves implementing a recognition system to successfully classify future samples, which are likely to be different from the training samples. Generalization is the ability of a classifier to deliver good performance on a test set of samples, which were not used during the training stage. Poor generalization may be attributed to any one of the following factors (De Baker 2002):

1. The number of features is too large relative to the number of training samples.
2. The number of unknown parameters associated with the classifier is large.
3. A classifier is too intensively optimized on the training set.

When performing classification of high-dimensional data such as images, the curse of dimensionality becomes an evident factor affecting classifier performance. The curse of dimensionality is a result of the sparseness of high-dimensional spaces. This implies that in the absence of simplifying assumptions, the amount of training data needed to get reasonably low variance estimators is fairly high. If the class densities are completely known, an increase in the number of features will not result in an increase in the probability of misclassification. However, it has been often observed in practice that the added features may actually degrade the performance of a classifier if the number of training samples that are used to design the classifier is small relative to the number of features. A simple explanation for this phenomenon is as follows (De Baker 2002).

As most classifiers use parametric estimation to find the probability density of the different classes, they estimate the unknown parameters and plug them in for the true parameters. For a fixed sample size, as the number of features is increased, the reliability of the parameter estimates decreases. Consequently, the performance of the resulting plug-in classifiers, for a fixed sample size, may degrade with an increase in the number of

features. Assuming that an important structure in the data actually resides in a smaller dimensional space is one method to handle this problem. Under this assumption, the goal is to reduce the dimensionality before attempting the classification. This approach can be successful if the dimensionality reduction/feature extraction method loses as little relevant information as possible in the transformation from high-dimensional space to the low-dimensional one.

4.1.5 Non-linear Features

Representing image information in the best way is a common objective in image recognition. Developing a good representation is key in exposing the constraints and removing the redundancies contained in pixel images. PCA-based schemes simplify the description of a set of interrelated data by representing data in terms of statistically independent variable or principle components. Here, each component (also referred to as an eigenvector or eigenimage) is a linear combination of the original variables. Upon choosing eigenvectors, any image in the set of training images can be approximately reconstructed with a linear combination of eigenimages, and their components will be stored in memory. Performing recognition of previously unseen images is accomplished by matching the image eigenvalues to those of the training set images. While PCA-based image recognition systems have been shown to be successful (Valentin *et al*. 1997), principal components are defined solely by their mean and standard deviation, and higher-order statistics are needed to form a good characterization of non-Gaussian data (Karhunen and Juotsensalo 1995). The problem is that the image subspace might be non-linear (e.g., face image subspace), which makes linear feature extraction methods, specifically PCA, a poor choice for extracting features from the data subspace. Due to the fact that PCA is mainly based on linear transformations, it is inappropriate for

modeling non-linear deformations such as bending. This is not the case for the DCT, as it can accommodate non-linear features quite well.

Several generalizations of PCA to deal with the problem of non-linear decomposition of random vectors into statistically independent components have been proposed (e.g., Karhunen and Juotsensalo 1995, Hyvarinen 1999). However, in general, PCA is data dependent and obtaining the principal components is a non-trivial task. Specifically, for face recognition, tremendous memory has to be used to store the data and an exhaustive search should be carried out to look for the closest match to an unknown face.

4.1.6 Motivation for DCT-based Features in the NoN Approach

Recall that KLT-based (PCA-based) transforms are optimal in the sense of energy compact and data deccorelation. However, the KLT is data dependent and computationally expensive transform. As this is the case, in first level of the proposed NoN model for face recognition, we generate DCT-based feature vectors, rather than KLT-based. The basis functions of DCT are data independent, and its information packing ability closely approximates the optimal KLT (Dinstein *et al*. 1990); the generation of DCT-based feature vectors provides a good compromise between information packing ability and computational complexity. In addition to this, the DCT has been shown to pack the most information into the fewest coefficients for most natural images and minimize *blocking artifact*[10]. Finally, in contrast to several other transforms, the DCT has already been implemented in a single integrated circuit due to its input independency.

[10] The block-like appearance that results when the boundaries between subimages become visible.

4.2 Network of Networks Hierarchical Clustering

When considering classification, a feature vector in a corresponding feature space is considered as "learned" when it is correctly clustered, or categorized. This clustering can be supervised or unsupervised in nature. The NoN model (see Figure 4.3, adapted from Sutton and Trainor 1991) also displays the capability for addressable memory in feature spaces. Following the development of a generalized hierarchical cluster model

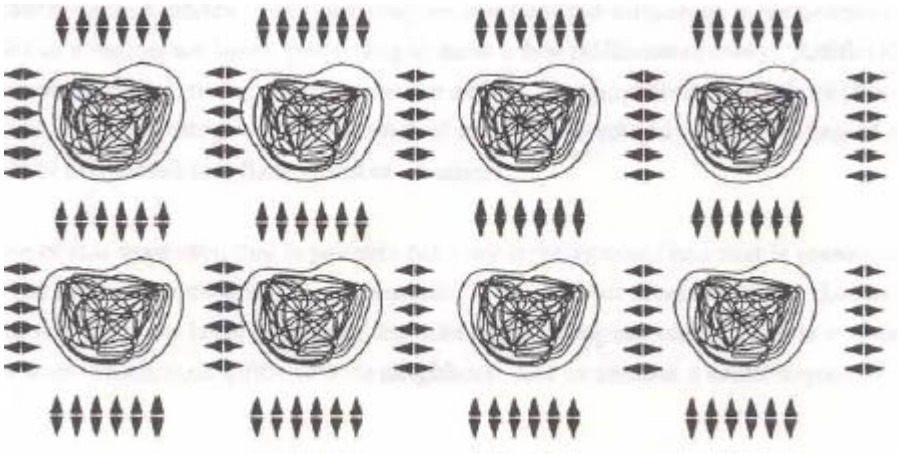


Figure 4.3: A network comprised of locally connected networks.

{Sutton *et al*., 1988), for illustrative purposes, a network consisting of three levels of neural organization was examined. The zeroeth level of organization consists of 3*N* binary state elements. These are arranged at the first level into three similar subunits, indexed by k = 1, 2, 3, and shown in Figure 4.4. Each subunit contains N elements with associated state values of $S_{ik} = \pm 1,\ i = 1,...,N$, the value S_{ik} characterizing the firing (+1) and the resting (-1) states of the i^{th} element in the k^{th} subunit. Within each subunit, p patterns of memory are embedded. These form the first level memory configurations. Pairwise connections exist between all the elements within a subunit (Sutton and Trainor 1991). Each subunit has the capacity for content addressable memory retrieval,

somewhat similar to the Hopfield (1982) model. The first level memory configurations in the cluster are somewhat analogous to the attractors in the Hopfield model.

Clustering among the three separate first level subunits is achieved by relatively sparse connections among the elements. These connections are, in general, asymmetric. They are generated by a collection of elements in each subunit that forms linkages with all the elements in their own subunit, as well as the elements in other subunits.

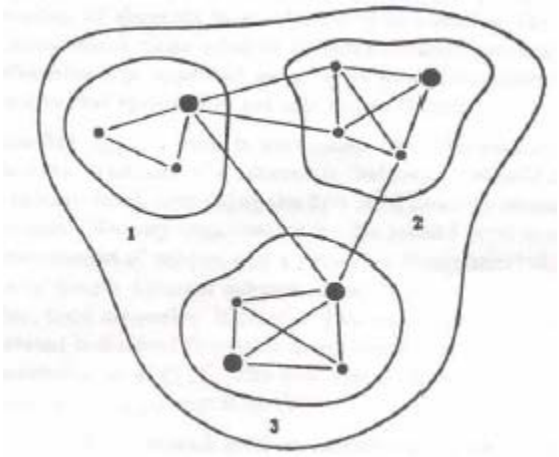


Figure 4.4: Subunit interconnectivity. Adapted from Sutton and Trainor (1991).

In Figure 4.4, three subunits 1, 2, 3 are depicted, each one containing *N* binary state elements. Some elements (small dots) connect to all other elements within a subunit, while other elements (large dots) connect to all the elements in their own and in other subunits. Only some of the latter type of connections are graphed for illustrative purposes (Sutton and Trainor 1991). The inter-subunit linkages always involve at least one large dot element, and the system has two levels of memory organization. Each subunit has stored patterns of element states, these being the first level memories. Some triplets of first level memories form memory patterns at a second level.

The network in Figure 4.4 can be viewed as having two populations of elements with different connectivities. In this context, one population form connections with every

element within its subunit and receives connections from within and between the subunits. The other population forms connections with every element in the entire network and receives connections from within and between the subunits. This arrangement attempts to capture the notion on neuronal heterogeneity, whereby some elements have only local connections, while other elements have both local and projective branchings (Sutton and Trainor 1991).

4.3 Summary

In this chapter we discussed features and their vector representation, as well as related topics such as feature spaces, feature extraction/selection and dimension minimization. Also discussed was the motivation for features based on the DCT, by the method proposed in this dissertation. The capability for addressable memory in feature spaces by the NoN model was reviewed as well. In Chapter 5, we present the proposed NoN system for face recognition, and give results obtained by this method on the AT&T database.

CHAPTER 5. TWO-LEVEL DCT COEFFICIENT BASED NoN SYSTEM

5.1 Introduction

In this chapter, we present a novel method of autonomic face recognition based on the recently proposed biologically plausible NoN model of information processing. In the general NoN case, an *n*-level hierarchy of nested distributed networks is constructed; in the proposed system we take *n* to be 2. The proposed system processes input face images by a nested family of locally operating networks along with a hierarchically superior network that classifies the information from each of the local networks.

In the initial discussion, we provide the system description for the NoN face recognition system, and subsequently give the specifics of operations at each level. We also introduce block-level DCT coefficient features, which are used by the hierarchically superior classifier in making a recognition decision. In Section 5.7 we give simulation results on the AT&T Cambridge Laboratories face database for the proposed method.

5.2 NoN Face Recognition System Description

A general description of the proposed NoN face recognition process is as follows. First, an input face image is divided into $N \times N$ blocks to define the local regions of processing. Next, the $N \times N$ 2-D DCT is used to transform the data into the frequency domain. The transformed version of the face is supplied to Level 1. Level 1 is composed of a nested family of locally operating networks that calculate various functions of spatial frequency in the block, producing a block-level DCT coefficient, which represents a compacted form of the spatial information in the block. At this point, the image is now transformed into a variable length vector and fed to Level 2. Level 2 is a hierarchically

superior classifier that is trained with respect to the data set. The output of Level 2 is the classification decision.

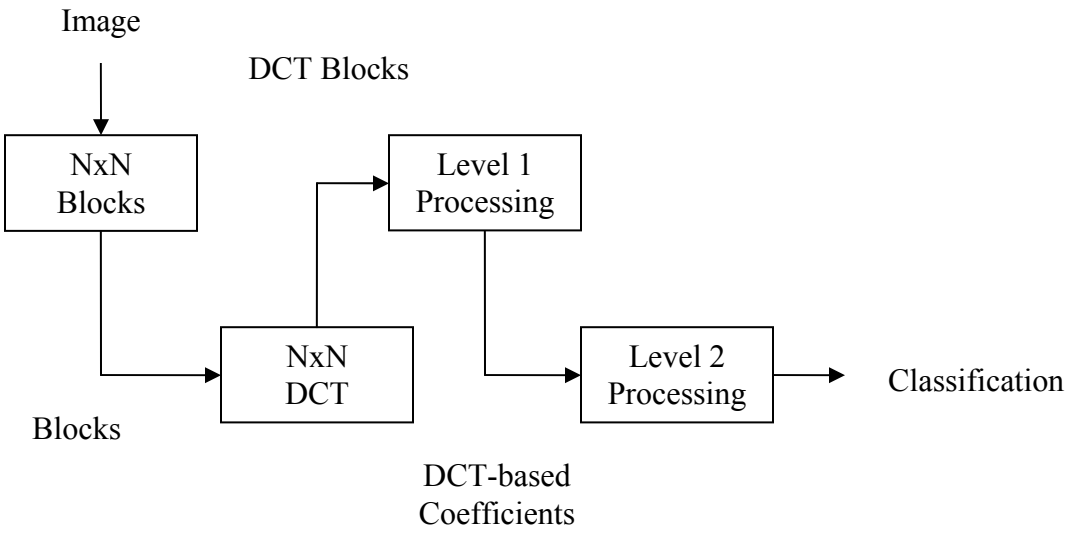


Figure 5.1: A block diagram description of the proposed system for face recognition.

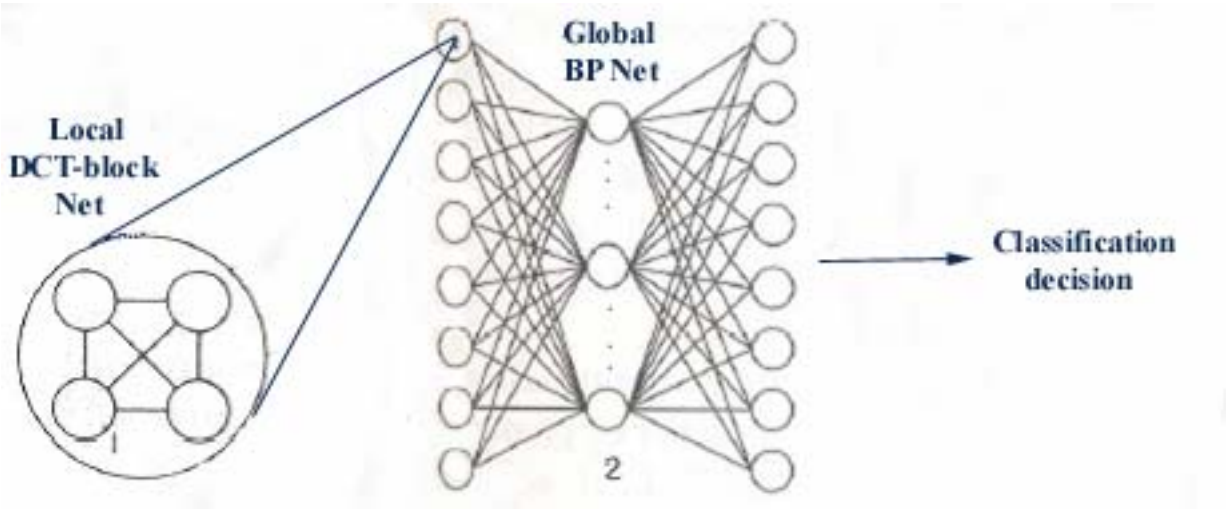


Figure 5.2: NoN model view of the face recognition system.

Given in Figures 5.1 and 5.2, are a block diagram description and a NoN model view of the proposed system, respectively.

5.3 DCT Use in Related Areas

In the proposed system, the DCT is used as the frequency transform of choice, due to the beneficial aspects of doing so, as discussed in Chapter 2. In recent literature, DCT has been used in machine vision applications with good results. In this section, we discuss two vision problems that employed the DCT.

5.3.1 DCT in Face Recognition

In Pan and Bolouri (1999), spatial frequency analysis using the DCT was proposed. The main idea of the approach is to compute the DCT coefficients of a face image, and select only a limited number of the coefficients (this limited selection corresponds to low spatial frequency coefficients) and use them as input to a multi-layer perceptron (MLP). In this method, the coefficient selection method is fixed in the system. The location of the transform coefficients retained for each image remains unchanged from one image to another.

The author takes the DCT of the entire image and selects a small number, ranging from 0.34% to 24.26% of the total number of coefficients (for a 92x112 input image, which was used in this work, this equals 35 – 2500 coefficients), starting at the top left corner of the DCT array. This is because low spatial frequency DCT components are mainly concentrated in the upper-left corner of the DCT array. The author states that this approach is reasonable because the reconstructed image from low frequency components is recognizable as a face. Once the low spatial frequency DCT coefficients are selected they are feed into to a multi-layer perceptron. In Pan *et al*. (2000), Pan and Bolouri (2000) and Pan and Bolouri (2001), the same method is followed, with only minor variation to the DCT coefficient selection process. The best mean error rate for

recognition testing this approach on the AT&T Cambridge Laboratories face database is given as (Pan and Bolouri 2001) 4.8% (see Table 3.1 for comparison).

5.3.2 DCT in Text Region Location

Applications like distance learning and teleconferencing often require compression of images that contain both text and graphics. Because text and graphics have different properties, a compression scheme can benefit by treating the textual and graphical portions of such compound images separately. Recently, methods for separating the textual and graphical portions of a compound image, namely, Transform Coefficient Likelihood (TCL) schemes, have been proposed (Keslassy *et al*. 2001). TCL schemes examine the DCT coefficient values of an 8x8 block. For each coefficient, they refer to stored histograms that give the likelihood that a certain value occurs in a text block, or in a graphics block. They then examine the differences in these two likelihoods over all the coefficients in the block to decide whether it contains text or graphics.

The DCT achieves compression in continuous tone images by exploiting the correlation among neighboring pixels. Text has different properties, however, and when it is compressed using a scheme developed for images, the compressed text will generally be at a lower quality for a given rate than surrounding graphics. It is beneficial, therefore, to identify regions in an image that contain text and treat them differently. Simply put, TCL schemes are methods to locate regions of text in an image.

To facilitate this comparison, implemented are functions that take as an input an 8x8 block of pixels or corresponding transform coefficients, and returns a real number known as the activity, a. If the activity is above some threshold, the 8x8 block is marked as text. TCL makes a decision $D_s(a(B))$ as:

$$D_s(a(B)) = \begin{cases} text & if \quad a(B) > t \\ graphics & \text{otherwise} \end{cases}$$

where $D_s(a(B))$ is the final decision, $a(B)$ is the activity of block B, and t is a specified threshold.

While TCL schemes do use the concept of block DCT computations, there are distinct differences. Approaching and solving the text location problem is much different than approaching face recognition. In TCL, block activation values are computed, and if the activation value is greater than some defined threshold, a Boolean variable is set, declaring the block as containing text. In the proposed method for autonomic face recognition, feature vectors are comprised of a set of block-level DCT coefficient values, which are computed and retained for each block of the input image, and applied to a classifier.

5.4 Motivation for Block-level DCT Coefficients in the NoN System

In this section, we provide reasoning for DCT-based feature vectors in the NoN recognition system.

5.4.1 Analogy with Biological Processing of Faces by Humans

The proposed method is somewhat analogous to the HVS in that information spanning the entire spatial frequency spectrum is used in face recognition. This spatial frequency information is obtained by using the DCT as an image transform, which has been shown to be advantageous in comparison to other frequency transforms. In the HVS, information about a face is typically obtained as an after-thought. Generally, this "extracted" information is used in recognition with very little effort. For machine vision of faces, however, recognition is not as trivial. Perhaps one of the most important aspects of a machine recognition system is the feature vector representation. As has been

discussed, the dimensionality alone of a feature vector can lead to very low recognition. In this dissertation, we use DCT-based features in classification because the energy compaction capability of the DCT closely approximates the optimal KLT, without the drawback of being computationally expensive. The goal is to model the HVS by extracting as much information from a face as possible.

The process of robust feature vector generation is one of the greatest challenges in face recognition. A system, when given an input image, ideally generates features or shape descriptors that capture the essential traits of a face, that are insensitive to environmental or illumination changes. The goal is to convey the essence of a given face image, while eliminating unnecessary information that serves to increase the dimensionality of the observation, and worse, weaken the signal strength of the image. In generating dimensionally reduced feature vectors, feature selection and extraction are commonly employed, and usually the selection criteria is improvement in classification accuracy.

A known problem in performing face recognition is to determine what facial features constitute the most relevant dimensions of a face. In attempting to solve this problem, there has been controversy with respect to the relevance of various spatial-frequency bandwidths for face perception and processing. Several authors, among whom Harmom (1973), Ginsburg (1978) and recently Pan and Bolouri (2001), have indicated that the low-frequency spectrum is sufficient for the processing of faces and that high frequencies convey only redundant information. By contrast, Fiorentini *et al*. (1983) have suggested that the information conveyed by the high frequencies is not redundant and in fact is sufficient by itself to ensure face recognition. What seems to be most evident is that the discrete facial features, such as the eyes, the mouth, and the chin, are

likely relevant units on which perception is based since they are inherent components of the face and determine its individuality (Sergent 1986). However, the availability of these facial features is not essential to recognize a face, as suggested by our capacity to identify an individual even when the facial features are blurred beyond recognition. For example, let us examine Figure 5.3 and Figure 5.4 adapted from Sergent (1986). In Figure 5.4, the intensity of each square is the average intensity of the equivalent area in the original face, so that all fine details have been filtered out. While the facial features are blurred, the locality of the features within the frame of the front-view face seems to provide enough information for recognition, even though the facial features have no specificity.

In observing Figure 5.4, the low-frequency representation of Figure 5.3, note that much of the relevant facial information is contained in the low spatial-frequency spectrum. Namely, a fair amount of information concerning sex, age, and general shape and configuration of the face is available. The low-frequency representation provides sufficient information concerning facial configuration but does not permit a detailed

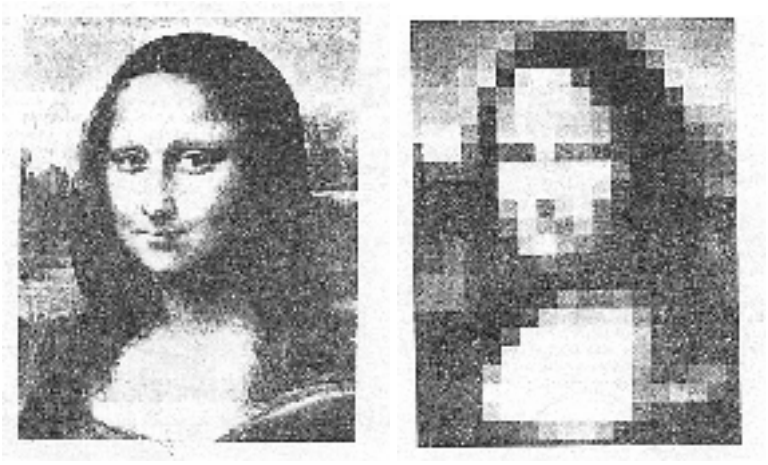


Figure 5.3 and 5.4, an original image of Mona Lisa and a low frequency image of Mona Lisa, respectively.

representation of the facial features. In other words, filtering out the high-spatial frequency contents makes the configuration of the face its main available property. Higher frequencies provide additional relevant information, particularly, that about the internal features and other details of the face, therefore, no particular range of spatial frequencies may be necessary as such for the perception and processing of faces (Sergent 1986).

It has been shown that human visual systems filter information contained in a display into separate bands of spatial frequencies, namely higher and lower frequency bands (De Valois and De Valois 1988). We can use the coarsely quantized picture of Mona Lisa (see Figure 5.4) as an analogy for some of the filtering taking place in the visual field. The retina is composed of a multitude of overlapping receptive fields of different sizes, such that each point of a visual scene is encoded several times as part of different channels. Large receptive fields are activated by the low frequency components of the picture while small receptive fields encode the high frequency components of the display (Sergent 1986).

5.4.2 Existing Efficient Algorithms for the DCT

Since its introduction, the DCT has found wide application in image and signal processing in general and in data compression in particular. It has been adopted as part of the standards for still and moving pictures coding. This is so, because it performs much like the statistically optimal Karhunen-Loeve transform under a variety of criteria.

Since the two dimensional discrete cosine transform (2-D DCT) is separable, it can be computed by the successive application of the 1-D DCT to the rows and columns. The formula for the 2-D DCT is separable, which means that it can be broken into two sequential 1-D DCT operations, one along the row vector and the second along the

column vector of the preceding row vector results. Called the Row-Column decomposition method, this is the most common method deployed for computing the 2-D DCT, and implementations usually focus on optimizing the 1-D DCT so that the Row-Column 2-D DCT implementation performs better when using the optimized 1-D DCT block along rows and columns. The 2-D DCT formula can be decomposed as follows:

$$F_{vu} = \frac{1}{4} C_v C_u \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} S_{yx} \cos\left(v\pi \frac{2y+1}{2N} \right) \cos\left(u\pi \frac{2x+1}{2N} \right)$$

$$F_{vu} = \frac{1}{2} C_v \sum_{y=0}^{N-1} \cos\left(v\pi \frac{2y+1}{2N} \right) \left[\frac{1}{2} C_u S_{yx} \cos\left(u\pi \frac{2x+1}{2N} \right) \right]$$

$$F_u = \frac{1}{2} C_u \sum_{x=0}^{N-1} S_x \cos\left(u\pi \frac{2x+1}{2N} \right)$$

By using the formula given above, instead of performing a 2-D DCT, we can perform a 1-D DCT on each row and then on each column of the block being processed. By computing the 2-D DCT as successive 1-D DCTs, performing computations will run faster than a single-pass 2-D DCT because we decompose a larger problem into a series of smaller (computationally less expensive) problems in order to compute the 2-D transform. The computational benefit of performing $N \; x \; N$ subimage (block) DCT calculations over direct 2-D DCT calculations is well known. Proposed in this dissertation is the use of block-level DCT coefficient features that are based on $N \; x \; N$ 2-D DCT coefficient calculations. This is because $N \; x \; N$ block DCT calculations can be implemented in-place requiring N^2 less memory locations and $2N^2$ less data transfers for the computation of $N \; x \; N$ DCT coefficients compared to existing direct algorithms. In practice, typically blocks of size 8x8 or 16x16 are used in computation (8x8 is optimum

for trade-off between compression efficiency and computational complexity). In subsequent sections, due to the computationally efficient properties of the DCT family of transforms, we employ the DCT-*II* as in JPEG, where N is even, i.e., $N = 2^m$, and $m > 2$.

5.5 Level 1: Nested Family of Locally Operating Networks

Level 1 of the proposed NoN system is a nested family of locally operating networks that calculate various functions of spatial frequency in a block, and we will show that these biologically motivated functions handle spatial information in a way analogous to the CSF. In this section, we begin with a general discussion of the necessary properties of an algorithm to be used in a block. Following this, the biologically motivated algorithms are presented.

5.5.1 Properties of Algorithms

An important property of an algorithm for computing block DCT coefficients is high-level energy compaction. In high-dimensional data, such as face images, the ability to access, store, and transmit information in an efficient manner is important. In a face recognition system, to utilize images effectively, actions must be taken to reduce the number of dimensions required for their representation. A general transform coding scheme involves subdividing an N x N image into smaller non-overlapping n x n sub-image blocks and performing a unitary transform (in our discussion, the DCT) on each block. The transform operation itself does not achieve any compression. It aims at decorrelating the original data and compacting the signal energy of the blocks into a relatively small set of transform coefficients. In general, a block-level DCT algorithm should further compact the spatial energy of the produced set of transform coefficients into a single representative block-level coefficient.

Perhaps the most essential property of a potential algorithm is the ability to operate on the spatial frequency information in a set of transform coefficients, in order to produce features that can be used for recognition in a manner analogous to the human vision system. As has been discussed, the human vision system combines the configuration cues of a face (lower spatial frequencies) with additional cues about the internal features and other details of the face (higher spatial frequencies) in a hierarchical manner, in performing face recognition. As a biologically motivated scheme, an algorithm for computing a block-level coefficient over a set of transform coefficients should also use cues from the entire spatial frequency spectrum in producing a representative coefficient.

5.5.2 Biologically Motivated Algorithms

In the proposed NoN system for face recognition, we apply the $N \times N$ 2-D DCT to all subimages (blocks) of a face image to reduce the information redundancy, and use a packed, vector form of the image in classification. To be more precise, let $C \in \Re^{NxN}$ be the 1-D vector representation, in raster scan order (see Figure 5.5), of the $N \times N$ 2-D DCT coefficients for a block. We define a function $\phi : \Re^{NxN} \rightarrow \Re$ as a function over the vector C. Ultimately, we present DCT-based feature vector representations of face images, based on varying the function ϕ, for use in classification. The effects of varying ϕ on the error rate in face recognition will also be evaluated.

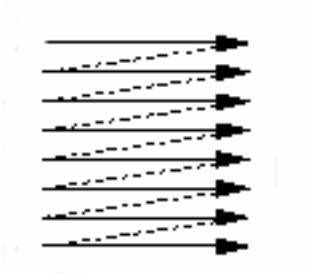


Figure 5.5: Raster scan order.

In all of the subsequent algorithms for computing block-level DCT coefficient features, the summation step for energy compaction over the DCT block begins at 2 in the 1-D raster scan order array. This is because the 2-D DCT places the DC (Direct Current) coefficient in the [0,0] position of the DCT array, and for frequency based calculations, luminance AC (Alternating Current) coefficients provide the most information concerning frequency information in the DCT block. Observing only the AC coefficients of the DCT block is reasonable, in that they convey the most useful information for use in spatial frequency energy compaction.

5.5.2.1 Direct Accumulation

The algorithms in this section accomplish high-level spatial energy compaction, while creating block-level DCT coefficient features that are biased towards the lower frequency spectrum. This biasing is motivated by the human capacity for face recognition of blurred or noisy images. As displayed by the low frequency image of Mona Lisa in Figure 5.4, a great deal of information is present in lower spatial frequencies. In fact, systems based solely on observation of these lower frequencies have performed reasonably well in recognition. Energy compaction and biasing towards the lower frequency spectrum is simultaneously accomplished by a summation of coefficient magnitudes over the DCT block. This is because low frequency coefficients are typically larger in magnitude than their high frequency counterparts.

The first scheme proposed for computing the block-level DCT coefficient, denoted as λ, is based on taking λ as the squared sum of the DCT coefficients c_i in the block, somewhat similar to activation calculations made in the text location problem using TCL schemes (Keslassy *et al*. 2001). Specifically, we let C be the set of DCT

coefficients of an $N\ x\ N$ block arranged in raster scan order. The block-level DCT coefficient λ for this scheme is then:

$$\lambda = \sum_{i=2}^{N^2} c_i^2 \qquad \text{(M1)}$$

where i = 2, …, N^2 such that $c_i \in C$, and having $N\ x\ N$ as the size of the DCT block.

A slight variation to the scheme in method (M1), replaces the squared term in the block-level DCT coefficient function with an absolute value, which is easier to compute (Chen 1990):

$$\lambda = \sum_{i=2}^{N^2} \left| c_i \right| \qquad \text{(M2)}$$

where i = 2, …, N^2 such that $c_i \in C$, and having $N\ x\ N$ as the size of the DCT block.

5.5.2.2 Direct Accumulation with Averaging

The algorithms introduced in this section are biologically motivated by the concept of the CSF, the transfer characteristic used to model the HVS as a measure of its response to spatial frequencies. Recall that in the HVS, receptive fields of the visual cortex display the highest sensitivity to the mid spatial frequency range. The algorithms in this section also perform high-level energy compaction, however, block coefficients generated by these methods can be considered as peaks in spatial energy corresponding to the middle of the spatial frequency spectrum. Biasing towards the middle of the frequency spectrum and energy compaction is handled by a summation of coefficient magnitudes over the DCT block, and then taking the average. This focus on the middle of the frequency spectrum is somewhat analogous to way spatial frequencies are biologically processed.

An algorithm for computing the block-level DCT coefficient λ, is based on taking λ as the squared sum of the DCT coefficients c_i in the block and dividing by the number of elements in the block. More specifically, we let C be the set of DCT coefficients of an $N \; x \; N$ block arranged in raster scan order. The block-level DCT coefficient λ for this scheme is then:

$$\lambda = \frac{\sum_{i=2}^{N^2} c_i^2}{N^2 - 1} \qquad \text{(M3)}$$

where i = 2, …, N^2 such that $c_i \in C$, and having N x N as the size of the DCT block. As in the direct accumulation scheme, a variation to method (M3) can be introduced to replace the squared term in the block-level DCT coefficient function with an absolute value.

$$\lambda = \frac{\sum_{i=2}^{N^2} \left| c_i \right|}{N^2 - 1} \qquad \text{(M4)}$$

where i = 2, …, N^2 such that $c_i \in C$, and having $N \; x \; N$ as the size of the DCT block.

5.5.2.3 Average Absolute Deviation

Another biologically motivated method for energy compaction and biasing towards the middle of the frequency spectrum is based on DCT block variance-based features. As in Section 5.4.2.3, the algorithm in this section attempts to model the CSF in processing spatial frequency information. The average absolute deviation of the luminance coefficients from the mean coefficient value in an DCT block is indicative of the overall energy in that block, which we claim to be useful for use in face recognition.

As a byproduct of the average absolute deviation computation, the block coefficient becomes partially invariant to variations (outliers) in energy of the block. The degree of invariance can be modified by adjusting the size of the block.

The proposed computation of block-level DCT coefficient takes λ to be the mean absolute deviation from the average value of each of the transform block elements c_i. Once again, we let C be the set of DCT coefficients of an N x N block arranged in raster scan order. The block-level DCT coefficient λ for this scheme is computed by:

$$\lambda = \frac{1}{N^2 - 1}\left(\sum_{i=2}^{N^2} \left| c_i - \mu \right| \right) \qquad \text{(M5)}$$

where μ is the mean of the DCT coefficients in the N x N block, computed by

$$\mu = \frac{\sum_{i=2}^{N^2} c_i}{N^2 - 1}$$

and i = 2, …, N^2 such that $c_i \in C$, where N x N is the size of the DCT block.

5.6 Level 2: Hierarchically Superior Backpropagation Network

In a general view of the proposed NoN system, Level 2 consists of a hierarchically superior network that classifies the information from each of the local networks. The specific network architecture chosen was the Backpropagation (BP) neural network model. In this section, we will briefly discuss the BP learning algorithm and related issues.

5.6.1 Backpropagation

The BP algorithm is well suited for expressing nonlinear decision surfaces, such as those corresponding to a face recognition task. Given a network with a fixed set of units and interconnections, the BP algorithm learns the weights for a multilayer network.

BP uses gradient descent with the goal of minimizing the squared error between the network output values and the target values for these outputs. An outline of the backpropagation algorithm is given by the following (Hertz *et al*.1991):

1. Randomly choose the initial network weights

2. While the stopping condition is not met

    2.1 For each pattern to be trained

    2.2 Apply the input feature vector to the network

    2.3 Calculate the output for every neuron from the input layer, through the hidden layer(s), to the output layer

    2.4 Calculate the error at the outputs

    2.5 Use the output error to compute error signals for pre-output layers

    2.6 Use the error signals to compute weight adjustments

    2.7 Apply the weight adjustments

    2.8 Periodically evaluate the network performance

The BP algorithm has been shown to work well in a variety neural network based applications. As this is the case, we BP learning is employed as the classification algorithm used by the Level 2 hierarchically superior network in the NoN model.

5.6.2 Feature Vector Normalization

Backpropagation theory suggests that we can speed up the training procedure if all input and output vectors are in the same range. In the proposed approach, inputs and outputs are scaled to lie in the [0, 1] range. In the network, the number of outputs is the number of face subjects (40 for the AT&T Cambridge Laboratories face database). To formulate this idea, suppose x_1, x_2,…,x_n are the computed block-level DCT coefficients

for a given image, where *n* is the number of $N\ x\ N$ blocks in the input image. The upper bounds (b_i) and lower bounds (a_i) can be determined by

$$b_i = \beta \cdot \max\{x_1, x_2, ..., x_n\} \quad i = 1, 2, ..., M;$$

and

$$a_i = \beta \cdot \min\{x_1, x_2, ..., x_n\} \quad i = 1, 2, ..., M;$$

Where *M* is the number of faces in the gallery and $\beta \geq 1$ is a factor to extend the bounds. Then an input vector $(z_1, z_2, ..., z_n)$ of the neural network is normalized by

$$z_i = \frac{x_i - a_i}{b_i - a_i} \quad i = 1, 2, ..., n.$$

For an unknown image, scaling factors are computed in the same manner as for training image, and input vectors normalized in this fashion are used as input vectors to the neural network.

5.7 Experiments and Discussion

In this section, a detailed discussion of NoN system results is provided. We begin our discussion with the stored representation of face images taken in this dissertation.

5.7.1 Stored Representation

Storing high-dimensional data, such as face images, in a space efficient way has become very important in recent times. Input images are fed to the system and divided into $N\ x\ N$ blocks, which are later used in $N\ x\ N$ DCT computations. In the proposed approach, input images are converted into 1-D arrays of block-level DCT coefficients and stored. Storing all face images in the gallery in this manner greatly reduces the space necessary to store images. This image representation helps to minimize storage cost, in that the dimension of the feature vector is greatly reduced, in comparison to pixel-based feature vectors.

In dividing a face image into N x N blocks, if the image dimensions do not divide by the subimage size, the image is zero-padded to the next multiple that will allow for division. Zero-padding is considered a pre-processing step to the proposed recognition algorithm. The following is an example of this zero-padding step used in this dissertation. The face images considered are those in the AT&T database, and all images are 92x112 (=10304 pixels) in dimension. Let us take N = 16 to be the desired block dimension, to produce 16x16 subimage blocks. The number of pixels in the subimage (256) does not divide the dimension of the image (10,304). However, if we scan the face image into a one-dimensional array (scanned in raster scan order) and zero-pad the image with 448 (now the face image has a total of 10,752 pixels) zeroes, the subimage divides the image 42 times. We can now determine reasonable common denominators of 42 as the number of 16x16 subimages of in the x and y direction. As an example, if we take 3 and 14, we can now divide the image into blocks of 16x16, because 3 x 14 x 256 = 10,752, which is the number pixels of the zero-padded face image. In simulations performed in this dissertation, N=8 required no zero-padding, N=16 required 448 zero-padded pixels, and N=32 required 1984 zero-padded pixels. Note that in the example given above, 42 equals the number of possible 16x16 subimages. Based on the block-level DCT algorithms, the total number of N x N subimages also corresponds to the number of block features required to represent the face image. By similar calculations to the example provided, it can be shown that the total number of features for N=8, N=16, and N=32, are equal to 161, 42, and 12, respectively. An example of the retained feature vectors for an example image (Figure 5.6) in the AT&T database is given below. In the example, the subimage blocks are of size 32x32, which yields 12 block-level DCT coefficients.

Figure 5.6: An example face image.

```
(M1):
[2189255.950   1903875.586    1576604.034  2249687.685
1902746.409    1138527.580    1813282.667  492808.238
2508068.414    1230060.172    498501.334   2612654.797]


(M2):
[9985.953  7688.156   8943.376   10005.956   7792.717
8849.055   15470.454  8878.777   16119.501   14398.697
9404.146   16892.072]


(M3):
[2140.035  1861.071  1541.157  2199.108  1859.967 1112.930
1772.515   481.728   2451.680  1202.405  487.294  2553.915]


(M4):
[9.761   7.515   8.742   9.781   7.618   8.650   15.123   8.679
15.757  14.075  9.193  16.512]


(M5):
[9.877   7.616   8.740   9.903   7.791   8.654   15.213   8.678
15.805  14.232  9.218  16.543]
```

5.7.2 Experimental Setup

In the conducted experiments, the weights of the Level 2 BP network are initialized to random values in [-1.0, 1.0]. The learning parameters, α the training rate, and μ, the momentum rate are both set to 0.3, by empirical determination. The maximum number of epochs is 300. The multiplication factor β in the equations for a_i and b_i is set to 1.0. The BP network consists of one hidden layer, with a variable number of neurons. For the AT&T Cambridge Laboratories face database, the number of outputs

of the neural network is always 40. To allow for comparisons, the same training and test set are used as in recent literature (as given in Chapter 3). Specifically, the first 5 images for each subject are the training images in total and the remaining 5 images are used for testing. As this is the case, there are 200 training images and 200 test images in total and no overlap exists between the training and test images. In each of the experiments, 5 random runs are carried out with randomly initialized weights for the BP network.

5.7.3 Experimental Results

Given in this section are the results obtained from 60 different system configurations. Specifically, we provide tabular and graphical results of recognition over all Level 1 energy compaction algorithms. The average error in recognition is computed over five runs of a given network configuration. The light gray denotes the minimum average error and the dark gray represents minimum classification error for a given method. If the minimum classification error and minimum average classification error were obtained due to a single network configuration, only one row in the table is highlighted. In the following, only resulted obtained using N = 8, 16, and 32 are provided. For values of N smaller than 8, the $N \times N$ DCT computations are not as efficient and the dimensionality of the feature vectors greatly increases, causing learning in the network to be extremely slow. For values of N greater than 32, the dimensionality of the feature vector becomes too small to convey enough information about the face for successful classification (e.g. if N = 64, the feature vector only has 3 elements).

Table 5.1: Recognition rates obtained when using method M1.

Method	NxN	No. of Coefficients	No. of Hidden	Avg. Error (%)	Min Error (%)
M1	8x8	161	15	14.2	11.0
	8x8	161	25	11.0	8.5
	8x8	**161**	**45**	**10.1**	**7.5**
	8x8	**161**	**60**	**9.0**	**8.0**
	16x16	42	15	17.5	14.5
	16x16	42	25	12.7	10.5
	16x16	42	45	11.1	9.0
	16x16	42	60	10.9	10.5
	32x32	12	15	14.5	12.5
	32x32	12	25	16.1	14.5
	32x32	12	45	16.4	14.0
	32x32	12	60	17.4	17.0

Table 5.2: Recognition rates obtained when using method M2.

Method	NxN	No. of Coefficients	No. of Hidden	Avg. Error (%)	Min Error (%)
M2	8x8	161	15	11.5	10.5
	8x8	161	25	8.0	6.0
	8x8	**161**	**45**	**4.0**	**4.5**
	8x8	**161**	**60**	**4.4**	**3.5**
	16x16	42	15	12.0	9.0
	16x16	42	25	8.9	7.0
	16x16	42	45	7.9	7.5
	16x16	42	60	6.9	4.5
	32x32	12	15	8.1	6.5
	32x32	12	25	10.1	7.5
	32x32	12	45	12.3	10.0
	32x32	12	60	12.8	10.5

Table 5.3: Recognition rates obtained when using method M3.

Method	NxN	No. of Coefficients	No. of Hidden	Avg. Error (%)	Min Error (%)
M3	8x8	161	15	14.3	13.5
	8x8	161	25	11.7	8.0
	8x8	161	45	9.9	9.5
	8x8	**161**	**60**	**8.7**	**7.0**
	16x16	42	15	14.1	12.5
	16x16	42	25	12.1	10.5
	16x16	42	45	13.2	11.5
	16x16	42	60	10.9	9.5
	32x32	12	15	14.3	13.0
	32x32	12	25	16.0	14.0
	32x32	12	45	16.2	14.0
	32x32	12	60	17.9	17.5

Table 5.4: Recognition rates obtained when using method M4.

Method	NxN	No. of Coefficients	No. of Hidden	Avg. Error (%)	Min Error (%)
M4	8x8	161	15	10.5	7.0
	8x8	161	25	9.0	7.0
	8x8	161	45	4.2	4.0
	8x8	**161**	**60**	**3.9**	**3.0**
	16x16	42	15	10.5	8.0
	16x16	42	25	9.0	8.0
	16x16	42	45	6.6	4.0
	16x16	42	60	7.1	4.5
	32x32	12	15	9.6	8.5
	32x32	12	25	9.8	7.0
	32x32	12	45	12.2	11.0
	32x32	12	60	14.4	11.5

Table 5.5: Recognition rates obtained when using method M5. Here † Denotes minimum error over all methods and network configurations.

Method	NxN	No. of Coefficients	No. of Hidden	Avg. Error (%)	Min Error (%)
M5	8x8	161	15	12.8	11.5
	8x8	161	25	6.0	2.5
	8x8	161	45	4.7	3.0
	8x8	**161**	**60**	**2.6**	**1.5 †**
	16x16	42	15	12.8	8.5
	16x16	42	25	9.2	7.0
	16x16	42	45	4.8	4.5
	16x16	42	60	7.9	6.0
	32x32	12	15	9.0	6.5
	32x32	12	25	9.9	8.5
	32x32	12	45	12.9	10.5
	32x32	12	60	12.1	11.0

The best results over all methods was obtained using $N = 8$, 60 hidden neurons, and utilizing method (M5). The minimum error and minimum average error were 1.5% (3 images incorrectly recognized of 200) and 2.6%, respectively. Given in Figure 5.7 is evolution of the recognition rate for the optimal network configuration. Figure 5.8 provides a plot of the best recognition results obtained employing each of the methods. As given in the Figure 5.8, method (M5), where DCT block coefficient are computed using the absolute average deviation, yielded the best highest recognition rates. Maximum recognition rates were obtained using 8x8 blocks over all methods. For methods (M2), (M3), (M4), and (M5), the system performed best with 60 hidden neurons. For method (M1), 45 hidden neurons produced the best results.

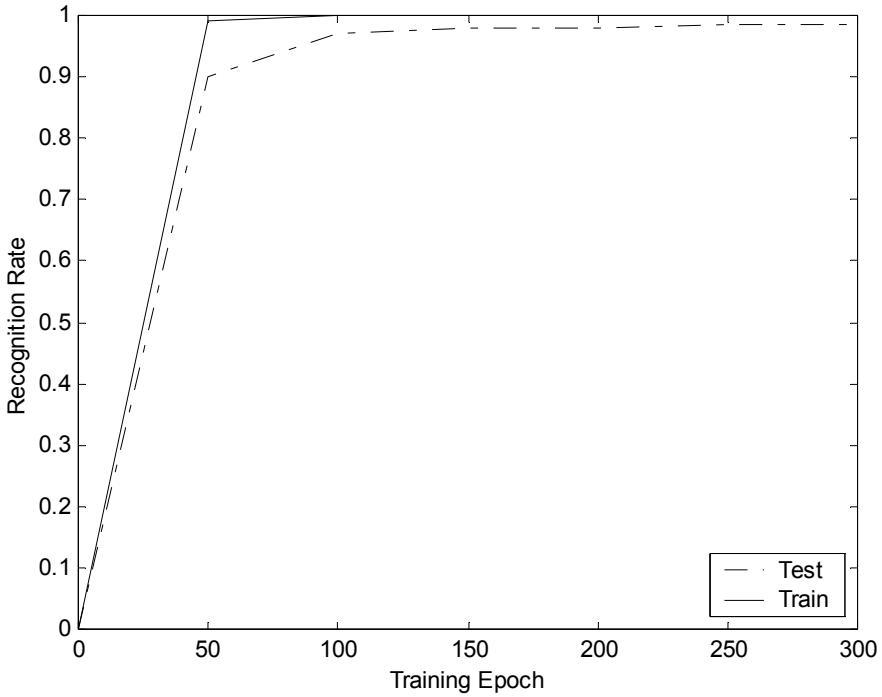


Figure 5.7: Evolution of recognition rate for 8x8 DCT block-level coefficients using M5 and 60 hidden layer neurons, which provided the maximum recognition rate of 98.5%.

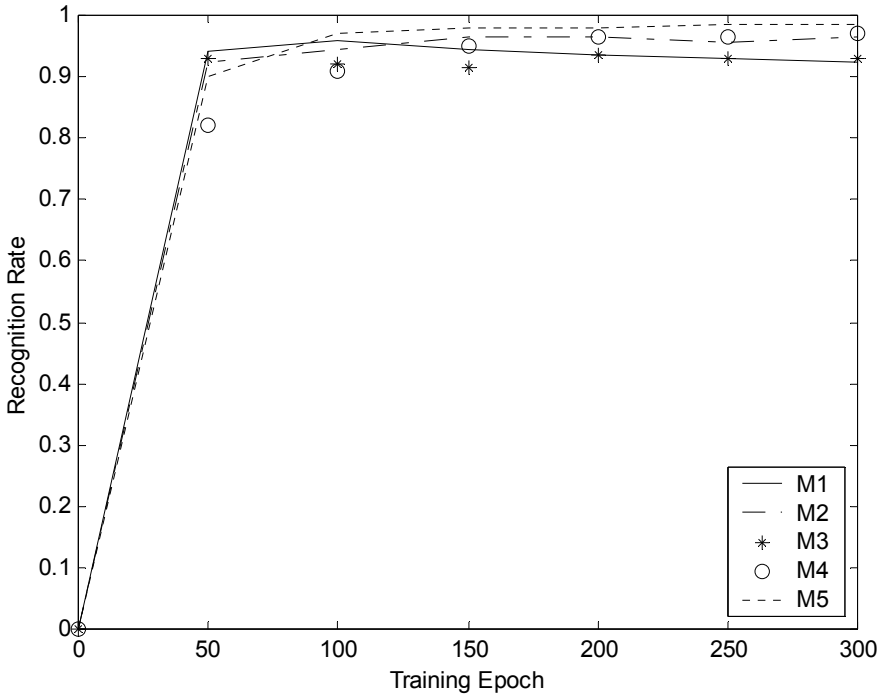


Figure 5.8: Plot of the best results obtained over all methods.

5.8 Drawbacks of Previous Biological Models

The proposed method for face recognition in this dissertation is motivated by biological processes of information processing and spatial vision. In this section we compare the presented biologically plausible NoN model for face recognition to other biological models of face recognition in recent literature. As is typically the case, we take the recognition rate attained by each model to be indicative of the prowess of the model.

5.8.1 Multiple Face Views Model

Samaria (1994) attempts to capture the human ability for face recognition based on recall of multiple views of a face. The drawback of this approach is recognition based on PCA, or Karhunen-Loeve expansion. Due to the data dependent nature of the KL expansion, a high computational cost is incurred by using this approach. The method proposed in this dissertation addresses this issue by employing the DCT which closely approximates KL expansion, while being competitive in terms of energy compaction and computational complexity.

5.8.2 Local Face Information Model

Hjelms (2000) attempts to capture the use of local features, in human face recognition, as the focal point of recognition. This is a drawback of the approach because one can view this as analogous to high-pass filtering of spatial frequencies. Local features of a face image include features such as the mouth, eyes, etc., and are very important in recognition. However, global information of face images is not used in classification. Consequently, this method displayed a relatively high error rate of 14.0%.

Table 5.6: Comparison of average error rates in current literature when using the AT&T Cambridge Laboratories face database.

System	Error %
Gaussian Weighting (Hjelms 2000)	14.0
Hidden Markov Models (Nefian and Hayes 1998)	14.0
Eigenface (Samaria 1994)	10.0
Linear Discriminant Analysis (Yu and Yang 2001)	9.2
Low frequency DCT with subimages (Pan and Bolouri 1999)	7.35
Low frequency DCT w/o subimages (Pan and Bolouri 2001)	4.85
Pseudo-2D Hidden Markov Models (Samaria 1994)	4.0
Probabilistic decision-based neural network (Lin *et al.* 1997)	4.0
Convolutional neural network (Lawrence *et al*.1997)	3.8
Linear support vector machines (Guo *et al*. 2001)	3.0
Block-level DCT Coefficient Features (proposed)	2.6
Kernel PCA (Kim *et al*. 2002)	2.5
One Spike Neural Network (Delorme and Thorpe 2001)	2.5
Uncorrelated Discriminant Transform (Zhong *et al*. 2001)	2.5

5.8.3 Global Face Information Model

In direct contrast to Hjelms (2000), Pan and Bolouri (1999), (2001) use holistic features of a face image and disregard local features. This is accomplished by retaining low spatial frequency DCT coefficients which correspond to global or configuration information about a face. In this work, the full-image 1-D DCT is applied to an input image, and a varying amount of lower frequency DCT coefficients are used as features. There are several drawbacks corresponding to this approach.

Firstly, as in Hjelms (2000), a substantial amount of information ignored by the feature extraction process. Information concerning local features of a face image is disregarded, and as has been shown by vision research, no single frequency sub-band is

sufficient for face recognition. The approach proposed in this dissertation addresses this issue by using global and local information in feature extraction; as a result, better classification results are reported.

A second drawback of this approach is the possibility for very high-dimensional feature vectors. On images in the AT&T database, the range of retained coefficients was between 35 and 2,500. It has been shown that high dimensional feature vectors typically degrade the classification accuracy of a classifier in sparsely sampled spaces. In addressing the issue of high-dimensional feature vectors, the NoN model employs high-level energy compaction via block-level DCT coefficient features. This energy compaction over a block greatly reduces the number of features necessary to represent a face image. In the proposed approach, on images in the AT&T database, the range of retained coefficients is between 12 and 161, as has been discussed previously.

Another drawback of Pan and Bolouri (2001) is the application of the full-image 1-D DCT to an input image. The 1-D full-image DCT has been shown to be much more computationally expensive than its 2-D DCT counterpart, due to existing algorithms for fast N x N 2-D DCT computation, as displayed in Section 5.3.2. The method proposed in this dissertation applies the computationally efficient N x N 2-D DCT as its basis image transform.

5.8.4 Visual Information Propagation Model

Delorme and Thorpe (2001) use the one spike neural network to model the manner in which visual information is propagated through the human vision system. In terms of the recognition rate, this model proved to be most robust. The drawback of this approach is the high computational cost incurred in training of the dedicated network. The network is organized into three layers of containing integrate-and-fire neurons. The

input layer included ON and OFF center cells whose activation levels depended on the local contrast at a given location in the input image. In the second layer of the model, orientation selectivity is performed using an array of 8 Gabor wavelets corresponding to 8 different orientations of a face image separated by 45°. In the output layer, the number of nodes corresponded to the number of individuals presented to the network. Training of the one spike network is computationally much more costly than the approach proposed in this dissertation due to the wavelet computations performed in the second layer of the model.

5.9 Summary

In this chapter, we have presented the two-level DCT coefficient based NoN system for face recognition. The system was given in diagrammatic form, in addition to specifics of simulation and experimentation. Algorithms for computing block-level DCT coefficient features and results based on employing each method were provided. Drawbacks of existing biologically motivated models for face recognition were also discussed. The proposed biologically plausible NoN model for face recognition has shown to be a robust model, attaining very good recognition rates.

CHAPTER 6. PARTITIONING SCHEMES AND ADAPTIVE BLOCKS

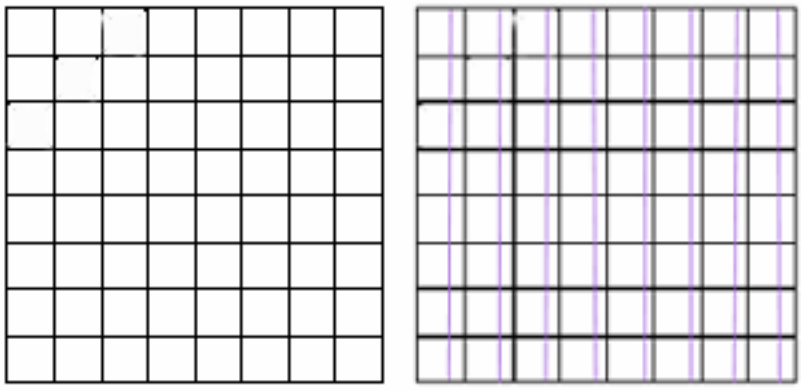
6.1 Introduction

Thus far, in the proposed NoN system, all input face images have been divided into $N \, x \, N$, non-overlapping blocks. In this chapter we discuss other possibilities for block partitioning of face images. Experiments based on overlapping blocks will also be discussed.

6.2 Block Partitioning

6.2.1 Overlapping Blocks

In the proposed approach, the original image is segmented into variable size blocks that contain homogeneous pixels. Secondly, the DCT is performed to each block. Following this, the set of coefficients in each block are applied to a statistical operator. The $N \, x \, N$ block partitioning performed is non-overlapping in nature (see Figure 6.1a).



a b

Figure 6.1: Partitioning schemes.

It is possible that in a non-overlapping scheme for image partitioning, that information can be missed at the boundaries of the blocks. An alternate scheme for partitioning is

based on sliding overlap of blocks (Figure 6.1b). When considering the sliding overlap case, information from a block, and its neighboring block, are used in computing block coefficients. In Section 6.3, we provide results of classification when considering this scheme for image partitioning.

6.2.2 Feedback in the NoN Model for Adaptive Blocks

In general, block partitioning need not be simply non-overlapping, or overlapping in nature. An interesting aspect of the NoN model is connectivity between networks and subnetworks. This connectivity can be viewed as a means for "communication" between the two hierarchical levels. As this is the case, we can envision a NoN based face

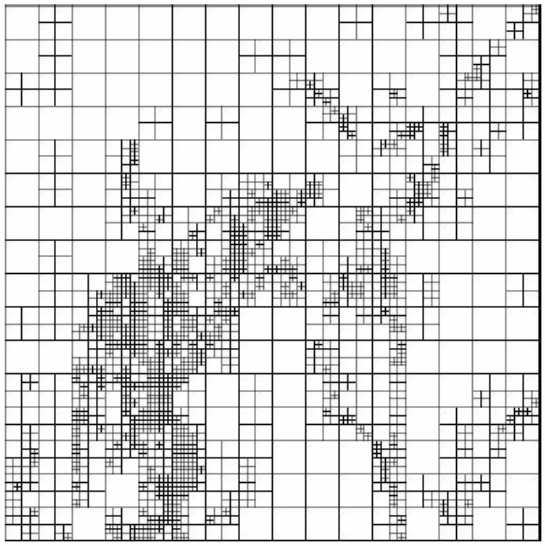


Figure 6.2: Adaptive Blocks

recognition model that incorporates feedback in both directions. This bi-directional communication implies that a decision made at the lower level can affect the higher level, and vice-versa. In terms of image partitioning, Level 2 of the NoN model could influence the way the images were partitioned based on attained global knowledge, and Level 1 based on local knowledge. In this context, certain areas of an input image that contain more information could be highly partitioned, and other areas of less interest could be

partitioned into larger blocks (see Figure 6.2) A partitioning scheme of this type would be adaptive in nature. While this is a futuristic view of the NoN model for face recognition, discussions of this nature are reasonable, and should be investigated.

6.3 Experiments and Results

Experiments have been performed, investigating the overlapping block case. In these experiments, we only reviewed system configurations that attained the best average results using the non-overlapping blocks case. As given in Chapter 5, the best average recognition rate (2.6%) for N=8 resulted with 60 hidden neurons, and applying algorithm (M5). For N=16, the best average recognition (4.8%) corresponded to algorithm (M5) and 45 hidden neurons. Finally, for N=32 best results we attained for algorithm (M2) and 15 hidden neurons (8.1%). By observing the best average recognition from the non-overlapping block case, we are able to determine if the system performance increases.

6.3.1 Experimental Setup

As in the experiments conducted in Chapter 5, the weights of the Level 2 BP network are initialized to random values in [-1.0, 1.0]. The learning parameters, α the training rate, and μ, the momentum rate are both set to 0.3, by empirical determination. The maximum number of epochs is 300. The multiplication factor β in the equations for a_i and b_i is set to 1.0. The BP network consists of one hidden layer, with a variable number of neurons. For the AT&T Cambridge Laboratories face database, the number of outputs of the neural network is always 40. The same testing and training set is used as in Chapter 5. In each of the experiments, 5 random runs are carried out with randomly initialized weights for the BP network.

6.3.2 Experimental Results

Given in Table 6.1 are results of experimentation using overlapping blocks. In the table, the average error in recognition is computed over five runs of a given network configuration.

Table 6.1: Error rates with overlapping blocks

NxN	Method	No. of Coefficients	No. of Hidden	Avg. Error (%)	Min Error (%)
8x8	M5	161	60	5.9	4.5
16x16	M5	42	45	9.8	7.5
32x32	M2	12	15	14.1	12.5

As displayed in the table, the best results were obtained using N=8, algorithm (M5) and 60 hidden neurons; this is consistent with the non-overlapping blocks case. However, overall, the recognition rates decreased when overlapping of blocks was allowed. This could be due to the introduction of redundant, rather than discriminatory, information by overlapping of the blocks. Perhaps a scheme for adaptive block partitioning, as suggested in Section 6.2, would serve to increase classification accuracy.

6.4 Summary

In this chapter non-overlapping, overlapping, and adaptive block partitioning schemes were discussed. Experiments were conducted using an overlapping block partition scheme. The obtained results displayed that non-overlapping blocks performed somewhat better in recognition. Adaptive block partitioning schemes were suggested as means to further improve upon classification results.

CHAPTER 7. CONCLUDING REMARKS

7.1 Introduction

The proposed NoN model for face recognition is motivated by biological information processing, namely cortical processing and spatial vision in the HVS. In comparison to other biologically inspired models, the NoN model performs quite well. The proposed approach has been shown to be a method of autonomic face recognition worthy of future study.

Recently the application of autonomic image recognition schemes to face recognition from still images has become an important area of research in present times. This is due to the numerous uses for such technology; commercial application and in the areas of security and law enforcement. For robust applications, reliable algorithms for recognition are required which work well regardless of conditions at image capture. The AT&T Cambridge Laboratories face database displays much of the variation typical in facial images, such as lighting, rotation, and scale. This is why the AT&T database has become popular in face recognition research as a benchmark database. Due to sufficient variation between images in the AT&T database, it is possible to extrapolate simulation results obtained on the database to larger image galleries. These are the reasons we chose to test the proposed method on this database.

7.2 Aspects of the NoN Model for Face Recognition

7.2.1 Analogy with Cortical Processing

The NoN model suggests that many neural circuits may be subdivided into essentially similar sub-circuits, where each sub circuit contains several types of neurons.

This hierarchy is true for the cerebral cortex of the human brain. Processing of visual stimuli is accomplished in area V1 of the cortex, and this processing is hierarchical. Due to its similar hierarchical nature, processing of face images using the NoN model is biologically plausible, it attempts to model human vision, and we expect it to be robust.

7.2.2 Analogy with Spatial Vision in the HVS

Another defining aspect of the NoN model is the use of spatial frequency information corresponding to a face in a manner analogous to the HVS. Spatial vision in the HVS utilizes information that spans the entire frequency spectrum. Cues about global characteristics of a face are present in lower spatial frequencies, while more specific information about discriminating features are present in higher frequencies. The combination of this information is used in a hierarchical fashion. The proposed approach performs recognition in a somewhat analogous way, by utilizing information from the entire spatial frequency spectrum.

The nested local networks of Level 1 of the NoN system handle spatial frequency information in a manner consistent with the CSF. The human vision system is most responsive to spatial frequencies in the middle of the spectrum, and the proposed method performs a somewhat analogous processing of spatial frequency information. When computing block-level DCT coefficients, by methods (M3), (M4) and (M5), feature vectors are produced that are biased towards the middle of the spatial frequency spectrum.

7.2.3 Drawbacks of Block-level DCT Features

While the proposed method exhibits promise as a model for biological processes in human face recognition, this approach suffers from certain limitations. One drawback is its susceptibility to blocking artifact degradation, due to the artificial discontinuities

that appear between the boundaries of the blocks, and these artifacts remain typically the greatest form of image degradation in block transform coding systems. Another drawback is the need to zero-pad images to make them divisible by the subimage blocks, and the introduction of this redundant data reduces classification accuracy. Yet another drawback of the approach is the computational cost incurred in computing the block-level coefficient features in Level 1 of the NoN model. While there is a computational benefit to applying the $N \times N$ 2-D DCT to an image, this benefit is reduced by performing energy compaction algorithms, based on summations, over the DCT block.

7.3 Future Research

7.3.1 Nonuniform Sampled Data Points

It may be useful to represent face images as nonuniformly sampled data points. In nonuniform sampling, the difference between two consecutive sample instances is dependent on the sampling conditions. Conditions at face image capture are seldom ever the same, and a robust face recognition system can account for these differences. Changes in illumination, background, facial features (hair, glasses, etc.) all affect machine recognition rates. Further research must be done in this area to develop methods to better represent signals corresponding to transformed images.

7.3.2 Alternate Level 1 Algorithms

The two important properties of the introduced NoN Level 1 algorithms are: (1) high-level energy compaction; and (2) the ability to operate on the spatial frequency information in a set of transform coefficients, to produce features that can be used for recognition in a manner analogous to the HVS. In our methods, biasing toward the lower and middle spatial frequency range was performed based on concepts of contrast sensitivity, as modeled by the CSF transfer characteristic.

Alternate Level 1 algorithms may be sought for computing block-level DCT coefficients features. Other functions may be based on logarithms, partial summations, and selectively summing certain DCT coefficients in a DCT block. The possibilities for alternate algorithms are countless in this context. Although the set of Level 1 algorithms introduced in this dissertation performed well, and the computations had a biological basis, recognition rates may improve using alternate algorithms.

7.3.3 Alternate Level 2 Hierarchically Superior Networks

The classifier used in this dissertation is a backpropagation (BP) neural network, but the popular BP classification has its drawbacks. BP neural networks, require iterative training, and sometimes never converge or take too long to train to be useful in real-time applications. The BP algorithm implements a gradient descent search through the space of possible network weights, iteratively reducing the error E between the training example target values and the network outputs. Because the error surface for multilayer networks may contain many different local minima, gradient descent can become trapped in any one of these. As a result, BP over multilayer networks is only guaranteed to converge toward some local minimum in E and not necessarily to the global minimum error. In addition to this, the BP algorithm is susceptible to overfitting the training examples at the cost of decreasing generalization accuracy over unseen examples. To address these problems, new classes of instantaneously trained neural network algorithms may be investigated (e.g., Kak 2002, Tang and Kak 2002).

7.3.4 Investigating Advanced NoN Models

The proposed 2-level NoN model has full connectivity as the higher level uses all the DCT-based coefficients simultaneously. It would be worthwhile to investigate systems where only spatially adjacent blocks are connected at the higher level.

Furthermore, three or higher-level systems may also be investigated. In a three-level system, the block size at the second level could be larger than the size at the first level. It would be useful to study the effect on performance of a system where the block sizes are adaptively adjusted in the learning phase.

.

REFERENCES

Abu-Mostafa, Y.S. and Psaltis, D. 1988. Optical neural computers. *Scientific American*, 256, pp. 88-95.

Ahmed, N. Natarajan, T. and Rao, K. 1974. Discrete Cosine Transform. *IEEE Trans. Computers*, 23, pp. 90-94.

Akamatsu, S. Sasaki, T. Fukamachi, H. and Suenaga, Y. 1991. A robust face identification scheme – KL expansion of an invariant feature space. *SPIE Proc.: Intell. Robots and Computer Vision X: Algorithms and Techn*., 1607, pp. 71-84.

Albrecht, D.G. De Valois, R.L. and Thorell, L.G. 1980. Visual cortical neurons: Are bars or gratings the optimal stimuli? *Science,* 207, pp. 88-90.

Anderson, J.A. Gately, M.T. Penz, P.A. and Collins, D.R. 1990. Radar signal categorization using a neural network. *Proceedings of IEEE,* 78, pp. 1646-1657.

Anderson, J.A. Spoehr, K.T. and Bennett, D.B. 1994. A study in numerical perversity: Teaching arithmetic to a neural network," *Neural Networks for Knowledge Representation and Inference Hills dale,* D. S. Levine and M. Aparicio Eds., NJ: Erlbaum.

Anderson, J.A. and Sutton, J.P. 1995. A network of networks: Computation and neurobiology. *World Congress of Neural Networks,* 1, pp. 561-568.

Apostolopoulos, J. and Jayant, N.S. 1997. Postprocessing for very low bit rate video compression. *IEEE Trans. on Image Processing*.

Banks, M.S. and Salapatek, P. 1981. Infant pattern vision: A new approach based on the contrast sensitivity function. *Journal of Experimental Child Psychology,* 31, pp. 145.

Banks, M.S. Stephens, B.R. and Hartmann, E.E. 1985. The development of basic mechanisms of pattern vision. Spatial frequency channels. *Journal of Experimental Child Psychology*, 40, pp. 501- 527.

Baron, R. 1981. Mechanisms of human facial recognition. *International Journal of Man Machine Studies*, pp. 137–178.

Bell, A. and Sejnowski, T. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7, pp. 1129–1159.

Bellman, R.E. 1961. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, Princeton, New Jersey, U.S.A.

Beymer, D.J. 1993. *Face Recognition under Varying Pose*. A.I. Memo No. 1461, Artificial Intelligence Laboratory, MIT.

Bledsoe, W.W. 1964. The model method in facial recognition. *Panoramic Research Inc., Tech. Rep. PRI:15*, Palo Alto, CA.

Blakemore, C. and Campbell, F.W. 1969. On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of physiology,* 203, pp. 237-260.

Blakemore, C.B. and Sutton, P. 1969. Size adaptation: a new aftereffect. *Science*, 166, pp. 245-247.

Blakemore, C.B. Carpenter, R.H.S. and Gerogeson, M.A. 1970. Lateral inhibition between orientation detectors in the human visual system. *Nature*, 228, pp. 37-39.

Braddick, O. Campbell, F.W. and Atkinson, J. 1978. Channels in vision: Basic aspects. *Handbook of sensory physiology,* 8, New York: Springer-Verlag.

Brunelli, R. and Poggio, T. 1992. HyperBF networks for gender classification. *Proc. DARPA Image Understanding Workshop*, pp. 311-314.

Buhmann, J. Lades, M. and Malsburg, C. 1990. Size and distortion invariant object recognition by hierarchical graph matching. *Proc., Int. Joint Conf. on Neural Networks*, pp. 411-416.

Campbell, F.W. and Robson, J. 1968. Application of Fourier analysis to the visibility of gratings. *Journal of Physiology,* 197, pp. 551-566.

Campbell, F.W. Cooper, G.F. and Enroth-Cugell, C. 1969. The spatial selectivity of the visual cells of the cat. *Journal of physiology,* 203, pp. 223-235.

Cannon, M.Y. 1979. Contrast sensation: a linear function of stimulus contrast. *Vision Res*., 19, pp. 1045-1052.

Canny, J. 1986. A computational approach to edge detection. *IEEE Trans. Patt. Anal. And Mach. Intell.*, 8, pp. 679-689.

Carl, J.W. Hall, C.F. 1972. The application of filtered transforms to the general classification problem. *IEE Trans. Comput*., 21, pp. 785-790.

Cardoso, J. 1998. Blind signal separation: statistical principles. *Proceedings of the IEEE. Special issue on blind identification and estimation*, 9, pp. 2009–2025.

Carey, S. Diamond, R. and Woods, B. 1980. The development of face recognition: a maturational component. *Developmental Psychology*, 16, pp. 257-269.

Chen, W. and Pratt, W. 1984. Scene adaptive coder. *IEEE Transactions on Communications*, 32, pp. 225-232.

Chen, C.T. 1990. Transform coding of digital images using variable block size DCT with adaptive thresholding and quantization. *SPIE,* 1349, pp.43-54.

Cheng, Y. Liu, K. Yang, J. Zhuang, Y. and Gu, N. 1991. Human face recognition method based on the statistical model of small sample size. *SPIE Proc.: Intell. Robots and Compu. Vision X: Algorithms and Techn.*, 1607, pp. 85-95.

Cheng, Y. Liu, K. Yang, J. and Wang, H. 1992. A robust algebraic method for human and face recognition. *Proc. 11th Int. Conf. on Patt. Recog*., pp. 221-224.

Chou, J. Crouse, M. and Ramchandran, K.1998. A Simple Algorithm for removing blocking artifacts in block-transform coded images. *IEEE Signal Processing Letters*, 5.

Comon, P. 1994. Independent component analysis, a new concept? *Signal Processing*, 36, pp. 287–314.

Cornsweet, T.N. 1970. *Visual perception.* New York: Academic Press.

Cortes, C. Krogh, A. and Hertz, J.A. 1987. Hierarchical associative networks. *Journal of Physics A*, 20, pp. 4449-4455.

Craw, I. Ellis, H. and Lishman, J. 1987. Automatic extraction of face features. *Patt. Recog. Lett.*, 5, pp. 183-187.

Craw, I. Tock, D. and Bennet, A. 1992. Finding face features. *Proc. 2nd Europe, Conf. on Computer Vision*, pp. 92-96.

Christopoulos, C. Bormans, L. Skodras, A. and Cornelius, J. 1994. Efficient computation of the two-dimensional fast cosine transform. *SPIE Hybrid Image and Signal Processing*, 5, pp. 229-237.

Daubechies, I. 1988. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math*., 41, pp. 909-996.

Daugman, J.G. 1984. Spatial Visual Channels in the Fourier Plane. *Vision Research*, 24, pp. 891-910.

De Baker, S. 2002. Unsupervised Pattern Recognition Dimensionality Reduction and Classification. *Ph.D. dissertation*, University of Antwerp.

Delorme, A. Gautrais, J. VanRullen, R. and Thorpe, S. J.1999. SpikeNET: A simulator for modeling large networks of integrate-and-fire neurons. *Neurocomputing*, 26, pp. 989-996.

Delorme, A. and Thorpe, S. 2001. Face identification using one spike per neuron: resistance to image degradations. *Neural Networks*, 14, pp. 795-803.

De Valois, R.L. De Valois, K.K. 1988. *Spatial Vision*. Oxford University Press. New York.

Devijver, P.A. and Kittler, J.V. 1982. *Pattern Recognition. A Statistical Approach*. Prentice-Hall, Englewood Cliffs, NJ.

Devroye, L. Gyorfi, L. and Lugosi, G. 1996. *A Problabilistic Theory of Pattern Recognition*. Springer-Verlag.

Diamantaras, K.I. and Kung, S.Y. 1996. *Principal Component Neural Networks: Theory and Applications*. John Wiley & Sons, Inc.

Dinstein, I. Rose, K. and Heiman, A. 1990. Variable block-size transform image coder. *IEEE Trans. on Communications*, November, pp. 2073-2078.

Duda, R., Hart, P. 1973. *Pattern Classification and Scene Analysis*. John Wiley, New York.

Ellis, H.D. 1986. Introduction to aspects of face processing: Ten questions in need of answer. *Aspects of Face Processing*, pp. 3–13.

Fiorentini, A. Maffei, L. and Sandini, G. 1983. The role of high spatial frequencies in face perception. *Perception*, 12, pp. 195-201.

Friedman, J. 1987. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82, pp. 246–266.

Fromherz, T. Stucki, P. and Bichsel, M. 1997. A Survey of Face Recognition. MML Technical Report, No 97.01, Dept. of Computer Science, University of Zurich.

Fu, K.S. 1982. *Syntactic Pattern Recognition and Applications*. Prentice Hall, New Jersey.

Fukunaga, K. 1989. *Statistical Pattern Recognition*. New York: Academic Press.

Fukunaga, K. 1990. *Introduction to Statistical Pattern Recognition*. Academic Press, New York.

Gabor, D. 1946. Theory of communication. *Journal of the Institute for Electrical Engineers*, 93, pp. 429–439.

Ginsburg, A.P. 1971. Psychological correlates of a model of the human visual system. *MS dissertation GE/EE/715-2*. Wright-Patterson AFB, Ohio: Air Force Institute of Technology.

Ginsburg, A.P. Carl, J.W. Kabrisky, M. Hall, C.F. and Gill, R.A. 1976. Psychological aspects of a model for the classification of visual images. *Advance in Cybernetics and Systems*, 3, pp. 1289-1306.

Ginsburg, A.1978. Visual Information Processing Based on Spatial Filters Constrained by Biological Data. *Ph.D. dissertation*, University of Cambridge.

Golomb, B.A. and Sejnowski, T.J. 1991. SEXNET: A neural network identifies sex from human faces. *Advances in Neural Information Processing Systems*, 3, pp. 572-577.

Goldstein, A.G. 1979a. Facial feature variation: Anthropometric data II. *Bulletin of the Psychonomic Society*, 13, pp. 191-193.

Goldstein, A.G. 1979b. Race-related variation of facial features: Anthropometric data I. *Bulletin of the Psychonomic Society*, 13, pp. 187-190.

Gopinath, R. Lang, M. Guo, H. and Odegard, J. 1994. Wavelet-based post-processing of low bit rate transform coded images. *Proc. of IEEE International Conference on Image Processing*.

Goshtasby, A. 1985. Description and discrimination of planar shapes using shape matrices. *IEEE Trans. Patt. Anal. And Mach. Intell*., 7, pp. 738-743.

Graham, N. and Nachmias, J. 1971. Detection of grating patterns containing two spatial frequencies: A comparison of single-channel and multiple-channel models. *Vision Research,* 11, pp. 251-259.

Graham, N. 1980. Spatial frequency channels in human vision: Detecting edges without edge detectors. *Visual Coding and Adaptability*, pp. 215-262.

Guan, L. 1994. Image restoration by a neural network with hierarchical cluster architecture. *J. Electronic Imaging,* 3, pp. 154-163.

Guan, L. Anderson, J.A. and Sutton, J.P.1997. A network of networks processing model for image regularization, *IEEE Transactions on Neural Networks*, 8, pp.169-174.

Guo, G.D. Li, S.Z. and Chan, K.L. 2001. Face recognition by support vector machines. *Proc. Int. Conf. Automatic Face and Gesture Recognition*, pp. 196–201.

Gutfreund, H.1988. Neural networks with hierarchically correlated patterns. *Phys. Rev.*, A, 37, pp. 570-577.

Govindaraju, V. Srihari, S.N. and Sher, D.B. 1990. A computational model for face location. *Proc. 3rd Int. Conf. on Computer Vision*, pp. 718-721.

Grenander, U. Chow, Y. and Keenan D. 1991. *Hands: A Pattern Theoretic Study of Biological Shapes*. New York: Springer-Verlag.

Grossman, A. and Morlet, J. 1984. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIMA J. Math.*, 15, pp. 723-736.

Hasan, M. Sharaf, A. and Marvasti, F. 1998. Subimage Error Concealment Techniques. *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS 98)*, 4, pp. 245-248.

Hasan, M. and Marvasti, F. 2001. Application of Nonuniform Sampling to Error Concealment. *Nonuniform Sampling: Theory and Practice*. Kluwer Academic/Plenum Publishers, New York.

Harmon, L.D. 1973. The recognition of faces. *Scientific American*, 229, pp. 71-82.

Hay, D.C. Young, A.W. 1982. The human face. *Normality and Pathology in Cognitive Functions*, pp. 173 – 202.

Heisele, B. Ho, P. and Poggio, T. 2001. *Face recognition with support vector machines: global versus component-based approach*. In Proc.8th International Conference on Computer Vision, Vancouver.

Henning, G.B. Hertyz, B.G. and Broadbent, D.E. 1975. Some experiments bearing on the hypothesis that the visual system analyses spatial patterns in independent bands of spatial frequency. *Vision Res*., 15, pp. 887-898.

Hertz, J. Krogh, A. and Palmer, R.G. 1991. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.

Heuther, B. 2002. The Visual Cortex [Internet] Available from: < http://www.geocities.com/b_huether/vcortex.html>

Hjelms, E. 2000. *Feature-Based Face Recognition*. Department of Informatics University of Oslo.

Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA,* 79, pp.2554-2558.

Hyvarinen, A. 1999. Survey on independent component analysis. *Neural Computing Surveys*, 2, pp. 94-128.

Hyvarinen, A. Karhunen, J. and Oja, E. 2001. *Independent Component Analysis*. John Wiley & Sons.

Intel Corp. 2002. Using Streaming SIMD Extensions in a Fast DCT Algorithm for MPEG Encoding [Internet] Available from: <http://www.intel.com/software/products/college/ia32/strmsimd/817down.htm>

Ioffe, L.B. and Feigerman, M.V. 1986. Asymmetry of interaction and hierarchy of patterns in the memory models. *ZhETF Lett*., 44, pp. 148-150.

Jain, A.K. and Farrokhnia, F. 1991. Unsupervised Texture Segmentation Using Gabor Filters. *Pattern Recognition*, 24, pp. 1167-1186.

Jin, Z. Yang, J. and Lu, J. 1999. An optimal set of uncorrelated discriminant features. *Chin. J. Comput.*, 22, pp. 1105-1108.

Kabrisky, M. Tallman, O. Day, C.M. Radoy, C. 1970. A theory of pattern perception based on human physiology. *Ergonomics*, 13, pp. 129-142.

Kak, S. 2002. A class of instantaneously trained neural networks. *Information Sciences*, 148, pp. 97-102.

Kanade, T. 1977. *Computer Recognition of Human Faces*. Basel and Stuttgart: Birkhauser.

Karhunen, K. 1947. Ueber lineare methoden in der wahrsheinlichtskeitsrechnung. *Annals Acad. Sci. Fennicae Series A. I*, 37.

Karhunen, J. and Joutsensalo, J. 1995. Generalization of principal component analysis, optimization problems and neural networks. *Neural Networks*, 8, pp. 549-562.

Kaya, Y. and Kobayashi, K. 1972. A basic study on human face recognition. *Frontiers of Pattern Recognition*, pp. 265-289.

Kelly, M.D. 1970. Visual identification of people by computer. *Tech. Rep. AI-130*, *Stanford AI Proj*.

Keslassy, I. Kalman, M. Wang, D. and Girod, B. 2001. Classification of Compound Images Based on Transform Coefficient Likelihood. *Proc. International Conference on Image Processing, ICIP-2001.*

Kim, K.I. Jung, K. and Kim, H.J. 2002. Face Recognition using Kernel Principal Component Analysis. *IEEE Signal Processing Letters* (SPL), 9, pp. 40-42.

Kohonen, T. 1988. *Self-Organization and Associative Memory*, Berlin: Springer.

Lades, M. Vorbruggen, J. Buhmann, J. Lange, J. Malsburg, C. and Wurtz, R. 1993. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers*, 42, pp. 300-311.

Lawrence, S. Giles, C. Tsoi, A. and Back, A. 1996. Face recognition: A hybrid neural network approach. Technical Report UMIACS-TR-96-16, University of Maryland.

Lawrence, S. Giles, C.L. Tsoi, A.C. and Back, A.D. 1997. Face recognition: A convolutional neural-network approach. *IEEE Trans. Neural Networks*, 8, pp. 98 113.

Lee, T.W. 1998. *Independent component analysis: Theory and Applications*. Kluwer Academic Publishers.

Le Gall, D.J. 1992. The MPEG Video Compression Algorithm. *Signal Processing: Image Communication*, 4, pp. 129-140.

Lin, S.H. Kung, S.Y. and Lin, L.J. 1997. Face recognition/detection by probabilistic decision based neural network. *IEEE Trans. Neural Networks*, 8, pp. 114–132.

Linares, I. Merserau, R. and Smith, M. 1994. Enhancement of block transform coded images using residual spectra adaptive postfiltering. *Proc. of IEEE Data Compression Conference*, pp. 321-330.

Liu, T. and Jayant, N.S. 1995. Adaptive postprocessing algorithms for low bit rate video signals. *IEEE Trans. on Image Processing*, 4, pp. 1032-1035.

Luo, J. Chen, C. and Parker, C. 1994. On the applications of Gibbs random field in image processing: from segmentation to enhancement. *SPIE Proc. of Visual Communications and Image Processing*, 2308, pp. 1289-1300.

Manjunath, B.S. Chellappa, R. and Malsburg, C.D. 1992. A feature based approach to face recognition. *Proc. IEEE Computer Soc.Conf. on Computer Vision and Patt. Recog*., pp. 373-378.

Marr, D. 1982. *Vision: A computational investigation into the human representation and processing* of *visual information.* San Francisco: Freeman.

Mountcastle, V.B. 1978. An organizing principle for cerebral function: The unit module and the distributed system. In *The Mindful Brain,* G. Edelman and V. B. Mountcastle Eds., pp. 7-50.

Movshon, J.A. Thompson, I.D. and Tolhurst, D.J. 1978. Spatial and temporal contrast sensitivity of neurones in areas 17 and 18 of the cat's visual cortex. *Journal of Physiology,* 283, pp. 101-120.

Nakamura, O. Mathur, S. and Minami, T. 1991. Indentification of human faces based on isodensity maps. *Patt. Recog*., 24, pp. 263-273.

Narendra, P. and Fukanaga, K. 1977. A branch and bound algorithm for feature subset selection. IEEE Trans. Computers, 26, pp. 917–922.

Nefian, A.V. and Hayes III, M.H. 1998. Hidden Markov Models for Face Recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2721-2724.

Netravali, A.N. and Haskell, B.G. 1988. Digital Pictures: Representation and Compression. *Applications of Communications Theory*. Plenum Press, NY, NY.

Pan, Z. Bolouri, H. 1999. High Speed Face Recognition Based on Discrete Cosine Transforms and Neural Networks. *Technical Report, Science and Technology Research Centre (STRC)*, University of Hertfordshire.

Pan, Z. Rust, G. and Bolouri, H. 2000. Image Redundancy Reduction for Neural Network Classification using Discrete Cosine Transforms. *Proc. of The IEEE-INNS-ENNS International Joint Conf. on Neural Networks (IJCNN2000)*, 3, pp. 149-154.

Pan, Z. Adams, R. Bolouri, H. 2000. Dimensionality Reduction of Face Images Using Discrete Cosine Transforms for Recognition. *Technical Report, Science and Technology Research Centre (STRC), University of Hertfordshire*.

Pan, Z. Adams, R. and Bolouri, H. 2001. Image Recognition Using Discrete Cosine Transforms as Dimensionality Reduction. *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP01)*.

Partridge, C.1993. *Gigabit Networking*, Addison-Wesley.

Pavlidis, T. 1997. *Structural Pattern Recognition*. Springer, Berlin, 2nd edition.

Pentland, A. Moghaddam, B. Starner, T. and Turk, M. 1994. View-based and modular eigenspaces for face recognition. *Proc. IEEE Computer Soc. Conf. on Computer Vision and Patt. Recognition.*, pp. 84-91.

Poggio, T. and Girosi, F. 1990. Network for approximation and learning. *Proc. IEEE*, 78, pp. 1481-1497.

Prabhakar, S. 2001. Fingerprint Classification and Matching Using a Filterbank. *Ph.D. Thesis*.

Rabiner, L. and Huang, B.1993. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ.

Rahardja, A. Sowmya, A. and Wilson, W. 1991. A neural network approach to component versus holistic recognition of facial expressions in images. SPIE *Proc.: Intell. Robots and Computer Vision X: Algorithms and Techn*., 1607, pp. 62-70.

Rao, K.R. and Yip, P. 1990. *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic Press, Boston.

Rao, K.R. and Hwang, J.J. 1996. *Techniques and Standards for Image, Video and Audio Coding*. Prentice-Hall, Upper Saddle River, NJ.

Regan, D. 1991. Spatial Vision. *Vision and Visual Dysfunction Series, no. 10.* MacMillan Press.

Reininger, R.C. and Gibson, J.D. 1993. Distributions of the two-dimensional DCT coefficients for images. *IEEE Transactions on Communications*, 6, pp. 835-839.

Sakai, T. Nagao, M. and Fujibayashi, S. 1969 Line extraction and pattern recognition in a photograph. *Pattern Recognition*, 1, pp. 233-248.

Samaria, F. and Young, S. 1994. HMM based architecture for face identification. *Image and Computer Vision*, 12, pp. 537-583.

Samaria, F.S. 1994. Face recognition using hidden Markov models. *Ph.D. dissertation*, Univ. Cambridge, Cambridge, U.K.

Seibert, M. and Waxman, A. 1991. Recognizing faces from their parts. *SPIE Proc.: Sensor Fusion IV: Control Paradigms and Data Structures*, 1611, pp. 129-140.

Seibert, M. and Waxman, A.M. 1993. An approach to face recognition using saliency maps and caricatures. *Proc. World Conf. on Neural Networks*, pp. 661-664.

Sergent, J. 1986. Microgenesis of face perception. *Aspects of Face Processing*, Dordrecht: Nijhoff.

Shepherd, J.W. 1985. An interactive computer system for retrieving faces. *Aspects of Face Processing*, Dordrecht: Nijhoff, pp. 398-409.

Sirohey, S.A. 1993. Human face segmentation and identification. *Tech. Rep. CAR-TR-695*, Center for Autom. Res., Univ. Maryland, College Park, MD.

Smoot, S.R. and Rowe, L.A. 1996. Study of DCT coefficient distributions. *Proceedings of the SPIE Symposium on Electronic Imaging*, 2657.

Stonham, T.J. 1984. Practical face recognition and verification with WISARD. *Aspects of Face Processing,* Dordrecht: Nijhoff, pp. 426-441.

Strang, G.T. 1989. Wavelets and dilation equations: A brief introduction. *SIAM Review*, 31, pp. 614-627.

Sutton, J.P. Beis, J.S. and Trainor, L.E.H. 1988. Hierarchical model of memory and memory loss. *J. Phys. A: Math. Gen.*, 21, pp. 4443-4454.

Sutton, J.P. and Trainor, L.E.H. 1991. Information Processing in Multi-levelled Neural Architectures. Proceedings Intern. AMSE Conference. *Neural Networks*, 1, pp. 59-66.

Sutton, J.P. and Anderson, J.A. 1995. Computational and neurobiological features of a network of networks. In *Neurobiology of Computation,* J. M. Bower Ed. Boston: Kluwer Academic, pp. 317 -322.

Szentagothai, J. 1977. The neuron network of the cerebral cortex. *Proc. R. Soc. Lond. B.,* 201, pp. 219-248.

Tang, K.W. and Kak, S. 2002. Fast Classification Networks for Signal Processing. *Circuits Systems Signal Processing*, 2, pp. 207-224.

Toulouse, G. Dehaene, S. and Changeux, J.P. 1986 Networks of formal neurons and memory palimpsests. *Proc. Natl Acad. Sci.,* 83, pp.1695-1698.

Turk, M.A. and Pentland, A.P. 1991. Face recognition using eigenfaces. *Proc. Int. Conf. on Patt. Recog.*, pp. 586-591.

Uhr, L.Vossler, C. and Uleman, J. 1962. Pattern recognition over distortions, by human still subjects and by a computer simulation of a model for human form perception. *Journal of Experimental Psychology,* 63, pp. 227-234.

Vaadia, E. Haalman, I. Abeles, M. Bergman, H. Prut, Y. Slovin, H. and Aertsen, A. 1995. Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. *Nature,* 373, pp. 515-518.

Valentin, D. Abdi, H. Edelman, B. and O'Toole, A. 1997. Principal component and neural network analyses of face images: What can be generalized in gender classification? Journal *of Mathematical Psychology*, 41, pp. 398-412.

Wallace. G.K.1990. *JPEG Technical Specification, Revision 5*, JPEG Joint Photographic Experts Group ISO/IEC JTC1/SC2/WG8 CCITT SGVIII, JPEG-8-R5.

Wallace, G. 1991. The JPEG still picture compression standard. *Communications of the ACM*, 34, pp. 30-44.

Wang, Z. and Hunt, B.R. 1985. The discrete W transform. *Applied Mathematics and Computation*, 16, pp. 1948.

Watson, A.B. 1993. Image compression using the Discrete Cosine Transform. *The Mathematica Journal*, 4, pp. 81-88.

Weng, J. and Swets, D.L. 1999. Face Recognition. *Biometrics: Personal Identification in Networked Society*, Boston, MA: Kluwer Academic, pp. 67-86.

Xiong, Z. Orchard, M.T. and Zhang, Y. 1996. A simple deblocking algorithm for JPEG compressed images using overcomplete wavelet representations. *IEEE Trans. on Circuits and Systems for Video Technology*.

Yang, G. and Huang, T.S. 1993. Human face detection in a scene. Proc. *IEEE Conf. on Computer Vision and Patt. Recog.*, pp. 453-458.

Yang, J. Yu, H. and Kunz, W. 2000. *An Efficient LDA Algorithm for Face Recognition*. ICARCV2000, Singapore.

Yang, Y. Galatsanos, N. and Katsaggelos, A. 1995. Projection-based spatially adaptive reconstruction of block-transform compressed images. *IEEE Trans. on Image Processing*, 4, pp. 896-908.

Yu, H. and Yang, J. 2001. A Direct LDA Algorithm for High-Dimensional Data -- with Application to Face Recognition. *Pattern Recognition*, 34, pp. 2067-2070.

Zhong, J. Yang, J. Hu, Z. and Lou, Z. 2001. Face recognition based on the uncorrelated discriminant transformation. *Pattern Recognition*, 34, pp. 1405-1416.

VITA

Willie L. Scott, a native of Louisiana, was born in 1977. He received his Bachelor of Science degree in December 1998 from Louisiana State University (LSU), majoring in computer engineering. Upon graduation, he continued his graduate studies at LSU in the area of computer engineering, and received his Master of Science degree in the summer of 2001. He furthered his graduate studies in the doctoral curriculum of the Engineering Science department at LSU, and will graduate with the degree of Doctor of Philosophy in that area.